

Retrieval-Augmented Multimodal Language Model with Heterogeneous Evidence Integration for Clinical Prognosis

Hamza Shafiq

Eindhoven University of Technology

m.h.shafiq@tue.nl

Jialu Tang

Eindhoven University of Technology

j.tang@tue.nl

Ting Dang

University of Melbourne

Ting.Dang@unimelb.edu.au

Jun Hu

Eindhoven University of Technology

j.hu@tue.nl

Aaqib Saeed

Eindhoven University of Technology

a.saeed@tue.nl

Abstract

Clinical prognosis requires integrating heterogeneous evidence from structured measurements, physiological waveforms, longitudinal history, and population context. Multimodal language models provide a natural framework for this integration, but standard approaches often serialize structured measurements as text, treat encounters independently, and lack a mechanism for incorporating representations of clinically similar patients. We introduce PRISM (Population-Retrieved Inference for Structured Multimodal Prognosis), which represents heterogeneous evidence through per-feature numerical tokenization, recency-weighted aggregation of longitudinal ECGs, and a dual-modality population memory bank. Retrieval operates in representation space: the model receives mean-pooled embeddings of nearest training-cohort encounters, while no individual retrieved record enters the prompt. On the 1,443-task MDS-ED benchmark (MIMIC-IV), PRISM attains the strongest reported figures across all four task groups: 86.2% Diagnosis, 93.1% Deterioration, 91.8% ICU admission, and 92.8% Mortality AUROC. These figures use a 96-hour pre-arrival window that the reported baselines did not receive. Under the benchmark input window alone, PRISM reaches 0.8895 AUROC, compared with 0.8937 for UniPACT and 0.8889 for MDS-ED. A matched-input compact fusion model receiving the same five evidence streams and frozen upstream encoders reaches 0.7652 AUROC, compared with 0.9092 for PRISM. Ablations show that replacing the numerical EHR pathway with full text serialization reduces AUROC by 3.66 points, removing the additional 96-hour pre-arrival history reduces it by 1.97 points, and zeroing retrieval reduces it by 0.99 points. Random-neighbor controls show a 0.97-point macro-average advantage for genuine retrieval; however, a paired task-level sign test does not establish a uniform task-level advantage ($p = 0.067$). On the external MC-MED cohort, supervised fine-tuning yields 0.8611 overall AUROC, compared with 0.8227 for locally trained gradient boosting and 0.7841 for source-only PRISM. These findings support heterogeneous evidence integration and supervised external adaptation.

1 Introduction

Clinical prognosis requires integrating multiple sources of patient information. Predicting whether a patient will deteriorate, require intensive care, or die in hospital demands the simultaneous interpretation of contin-

uous physiological waveforms, dense panels of laboratory analytes and vital signs, longitudinal acquisition histories, and the implicit epidemiological priors that experienced clinicians carry from thousands of prior cases (Smith et al., 2013; Raghunath et al., 2020). Large language models have emerged as a compelling backbone for this integration: their capacity for cross-modal attention over heterogeneous token sequences offers, in principle, a unified substrate for fusing structured and unstructured clinical data into coherent prognostic judgments (Singhal et al., 2023). In practice, however, current multimodal clinical LLMs leave substantial predictive signal unrecovered due to three structural limitations in how they represent, contextualize, and reason over clinical evidence.

The most immediate of these concerns how structured clinical data enters the model. Systems that fuse waveform encoders with language model backbones (Liu et al., 2023; Li et al., 2023; Tang et al., 2026; Pham et al., 2025) typically rely on *textualization*: laboratory values, vital signs, and demographics are serialized into natural-language strings and consumed through the tokenizer. This conversion does not explicitly preserve the ordinal and ratio-scale structure of numerical measurements at the input boundary. A troponin of 0.11 ng/mL and a troponin of 0.01 ng/mL become sequences of subword tokens whose clinical distance is opaque to the embedding layer (Gorishniy et al., 2021). Analytes such as troponin, lactate, and creatinine, among the strongest independent predictors of acute deterioration and in-hospital mortality (Ahmed et al., 2014; Kruse et al., 2011), are either collapsed into coarse text summaries or discarded entirely. The resulting token sequence therefore carries the values without an explicit encoding of their numerical geometry.

A related limitation concerns the temporal dimension. Serial changes in a patient’s ECG carry independent prognostic information beyond any single acquisition (Lehmacher et al., 2020): a new ST deviation against a previously normal baseline has different clinical meaning than the same morphology in a patient with chronic abnormalities. Yet current multimodal systems encode only the index-visit waveform (Tang et al., 2026; Zhao et al., 2025; Pham et al., 2025), treating each encounter as if it were the patient’s first and forfeiting the prognostic signal embedded in longitudinal change. This is a particularly costly omission given that ECG foundation models have shown that waveform representations encode risk information well beyond traditional diagnostic categories (Raghunath et al., 2020; Li et al., 2025; Tsai et al., 2023).

Finally, a language model without an explicit retrieval mechanism cannot consult representations of similar patients in the target population. Such representations may provide useful context beyond the current encounter. Retrieval-augmented generation has proven effective for grounding language models in external text corpora (Lewis et al., 2020) and has been applied to clinical NLP with knowledge graphs (Zhu et al., 2024). We study retrieval from structured patient populations, where the retrieved evidence consists of learned encounter representations.

We introduce **PRISM** (**P**opulation-**R**etrieved **I**nference for **S**tructured **M**ultimodal **P**rognosis)¹, a multi-modal architecture that supplies a language model with population-level context while representing structured clinical measurements explicitly. Rather than treating the language model as a self-contained reasoner that must absorb all clinical context through its tokenizer, PRISM equips it with three external capacities: a dense numerical encoder that represents every clinical feature as an individual learned token, a temporal aggregation mechanism that distills longitudinal waveform history into a compact representation under exponential recency weighting, and a population retrieval module that supplies representations of historically similar patients as non-parametric context. Each capacity is realized through a dedicated component: an FT-Transformer (Gorishniy et al., 2021) for structured EHR encoding, an exponential recency-weighting module for ECG history aggregation, and a dual-modality population memory bank over ECG and EHR embeddings. Their outputs are projected as soft tokens into a shared sequence consumed by a medical language model (i.e., MedGemma (Sellergren et al., 2025)) fine-tuned with low-rank adaptation (Hu et al., 2022), and a single trained model answers all 1,443 task questions, each in its own forward pass, without task-specific output heads. The mechanisms we build on are established individually, and per-feature numerical encoding in particular is standard practice in tabular and relational deep learning (Gorishniy et al., 2021). What we contribute is their composition into a single tokenized interface over heterogeneous clinical evidence, together with an empirical account of how performance changes as individual evidence streams are removed or replaced across task groups.

¹Code: <https://github.com/hamzashafiq28/PRISM>.

We evaluate PRISM on the MDS-ED benchmark (Lopez Alcaraz et al., 2025), the most comprehensive multimodal clinical prognosis benchmark available, comprising 1,443 binary prediction tasks over raw 12-lead ECG waveforms and 470 structured clinical variables from MIMIC-IV (Johnson et al., 2023). PRISM attains the strongest reported figures across all four task groups: 86.24% on Diagnosis (1,428 ICD-10 tasks), 93.12% on Deterioration (6 tasks), 91.79% on ICU admission (2 tasks), and 92.82% on Mortality (7 tasks). Counting tasks whose lower 95% confidence bound exceeds an AUROC of 0.80, the criterion used by UniPACT, PRISM qualifies on 1,145 of 1,443 tasks (79.3%) against 883 for UniPACT (61.2%) and 623 for MDS-ED (43.2%), gains of 262 and 522 tasks respectively. Controlled ablations quantify the corresponding changes: full text serialization in place of the numerical EHR pathway is -3.66 AUROC points, removing the pre-arrival window -1.97 , removing the current ECG -1.61 , removing longitudinal history -1.22 , and zeroing retrieval -0.99 . We additionally evaluate external transfer and supervised adaptation on MC-MED (Kansal et al., 2025), with controlled locally trained baselines. Fine-tuned PRISM has the highest overall point estimate in this external comparison, and supplementary temporal validation compares PRISM with compact fusion baselines and retrieval controls. Our principal contributions are:

- We propose PRISM, a multimodal architecture for clinical prognosis that unifies dense numerical encoding of structured clinical records, temporal aggregation of longitudinal waveform history, and population-level retrieval-augmented generation within a single language model reasoning framework. Against a 5.9M-parameter fusion model receiving the same five input streams and frozen encoders, PRISM improves AUROC by 14.40 points, and by 13.33 points against the strongest compact three-stream configuration. These gains compare complete model configurations under the controls described in Section 4.5.
- We show that replacing the compact numerical soft-token pathway with full text serialization of the same structured record reduces AUROC by 3.66 points, with the largest differences on diagnosis tasks that depend strongly on continuous-valued laboratory measurements.
- We introduce a temporal history aggregator that pools prior-visit ECG embeddings under exponential recency weighting. On the 122 encounters where pooling is operative, the learned rule has the highest point estimate among the tested alternatives: $+0.67$ AUROC points over the best fixed decay, $+1.10$ over uniform averaging, and up to $+2.33$ over single-acquisition selection.
- We extend retrieval-augmented generation from text corpora to empirical patient populations by indexing per-encounter ECG and EHR embeddings and injecting mean-pooled neighbor representations as additional soft-token context. Retrieval operates entirely in representation space, so the model gains population context while no individual retrieved record enters the prompt.

2 Related Work

Multimodal clinical LLMs and structured data encoding. The soft-token alignment paradigm, in which encoder outputs are projected into a language model’s embedding space as learnable tokens (Liu et al., 2023; Li et al., 2023), has become the dominant approach for multimodal clinical prediction. Recent systems apply this paradigm to align ECG waveform encoders (Li et al., 2025; Tian et al., 2024; Hung et al., 2025) with LLM backbones for diagnostic and prognostic tasks (Tang et al., 2026; Pham et al., 2025; Zhao et al., 2025). A persistent design choice across these systems is that structured clinical data, including laboratory values, vital signs, and demographics, enters the language model as serialized natural-language strings rather than through encoders that preserve numerical fidelity. Transformer architectures designed for heterogeneous tabular data, such as the FT-Transformer (Gorishniy et al., 2021), map each feature to an individual token via learned per-feature projections. This provides a numerical encoding mechanism for multimodal clinical prediction. Separately, pre-trained EHR Transformers such as Med-BERT (Rasmy et al., 2021) and BEHRT (Li et al., 2020) learn effective representations from longitudinal diagnostic code sequences but target discrete event vocabularies rather than the dense continuous-valued features that characterize acute-care data. We integrate an FT-Transformer based encoder directly into the multimodal LLM pipeline, encoding all 470 clinical features numerically and providing an explicit numerical representation alongside the textual prompt.

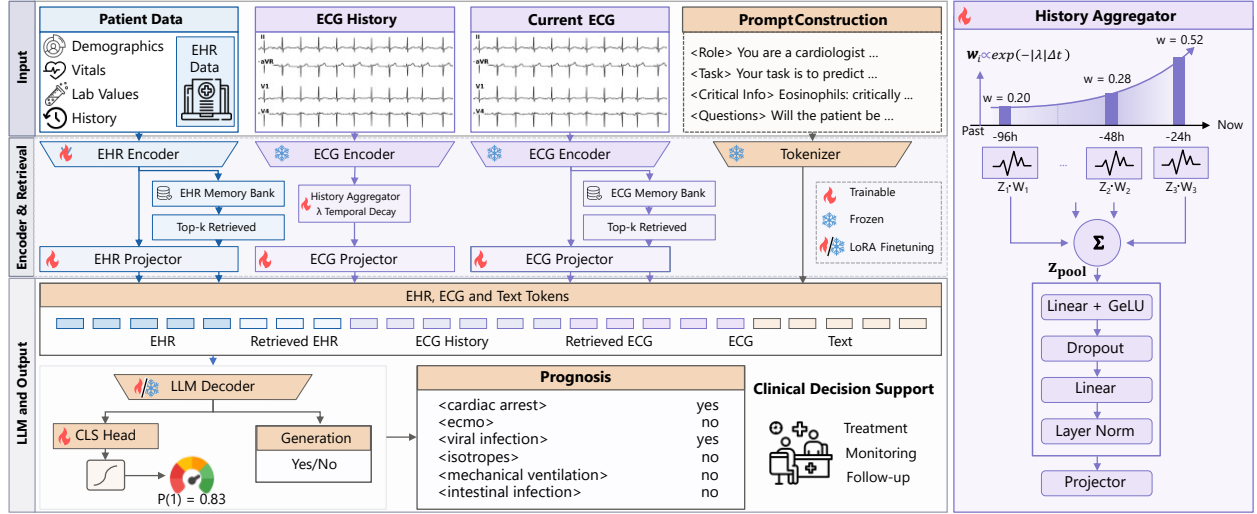


Figure 1: Overview of the PRISM architecture. Five modality streams (structured EHR, current ECG, longitudinal ECG history, retrieved EHR context, and retrieved ECG context) are independently encoded and projected as soft tokens into a shared sequence consumed by a LoRA-adapted MedGemma-4B decoder. The population memory bank (center) retrieves the $k=3$ nearest training-fold encounters from dual EHR and ECG indices constructed after Stage 1 and passes their mean-pooled embeddings, not their outcome labels, to the decoder. The temporal history aggregator (right) pools prior ECG embeddings under exponential recency weighting. Each task question is answered in one forward pass, via either the generative or the classification pathway, with no task-specific heads.

Longitudinal modeling in clinical prediction. Physiological signals acquired serially carry prognostic information beyond any single measurement. In the ECG domain, serial waveform evolution provides independent predictive value for adverse cardiac events in acute presentations (Lehmacher et al., 2020), and large-scale studies have confirmed that ECG waveforms encode risk patterns invisible to expert interpretation (Raghunath et al., 2020; Tsai et al., 2023). Existing multimodal clinical LLMs, however, encode only the index-visit acquisition, treating each encounter in isolation. Longitudinal EHR models (Rasmy et al., 2021; Li et al., 2020) capture temporal dependencies over months or years of discrete clinical events, but the problem of incorporating short-horizon, multi-acquisition waveform histories within a language model architecture remains unaddressed. We address this gap through a temporal history aggregator that pools prior-visit ECG embeddings via exponential recency weighting governed by a single scalar, injecting longitudinal waveform context into the language model at negligible parameter cost.

Retrieval-augmented reasoning in healthcare. Retrieval-augmented generation (or RAG) improves factual consistency by grounding language model outputs in external evidence (Lewis et al., 2020). In clinical settings, this motivation is particularly strong: even large language models encoding substantial medical knowledge exhibit factual inconsistencies under expert scrutiny (Singhal et al., 2023). EMERGE (Zhu et al., 2024) applies retrieval over a clinical knowledge graph to generate task-relevant patient summaries for mortality and readmission prediction. Clinical RAG commonly retrieves from knowledge sources such as guidelines, ontologies, or literature. Patient-population retrieval instead supplies per-encounter embeddings as context. We investigate this setting by indexing learned encounter representations from the training cohort in a dual-modality memory bank and supplying mean-pooled representations of clinically similar cases as additional context for the language model.

3 Method

3.1 Problem Formulation and Overview

Let each clinical encounter be characterized by a tuple $(\mathbf{X}_{\text{ehr}}, \mathbf{X}_{\text{ecg}}, \mathcal{H}, \mathbf{p})$, where $\mathbf{X}_{\text{ehr}} \in \mathbb{R}^D$ is a D -dimensional structured clinical feature vector (here $D=470$, comprising laboratory analytes, vital signs, demographics, and categorical indicators), $\mathbf{X}_{\text{ecg}} \in \mathbb{R}^{12 \times T}$ is a raw 12-lead ECG waveform of length T , $\mathcal{H} = \{(\mathbf{X}_{\text{ecg}}^{(i)}, \delta_i)\}_{i=1}^N$ is a set of $N \geq 0$ prior ECG acquisitions with corresponding temporal offsets $\delta_i > 0$ (in

hours before the index visit), and $\mathbf{p}$ is a natural-language clinical prompt specifying the prediction target. Given a set of $\mathcal{T}$ binary outcome tasks (here $|\mathcal{T}|=1,443$), the goal is to learn a function f that maps each encounter to a vector of probabilities $\hat{\mathbf{y}} = f(\mathbf{X}_{\text{ehr}}, \mathbf{X}_{\text{ecg}}, \mathcal{H}, \mathbf{p}) \in [0, 1]^{|\mathcal{T}|}$, where each $\hat{y}_t$ estimates $P(y_t=1)$ for outcome t . PRISM decomposes this function into five encoding streams, a population retrieval mechanism, and a shared language model decoder, as illustrated in Figure 1. Each stream produces a fixed-length embedding that is projected as a sequence of soft tokens into the language model’s embedding space. Prompt design and data construction are described in Appendices J and K.

3.2 Structured EHR Encoding

We encode the complete D -dimensional clinical feature vector using a Feature Tokenizer Transformer (FT-Transformer) (Gorishniy et al., 2021), which preserves the numerical fidelity of each clinical measurement by mapping it to an individual learned token rather than serializing it as text. Concretely, each continuous feature x_j is projected to a d_{model} -dimensional token via a learned per-feature affine transformation $\mathbf{t}_j = x_j \mathbf{w}_j + \mathbf{b}_j$, where $\mathbf{w}_j \in \mathbb{R}^{d_{\text{model}}}$ and $\mathbf{b}_j \in \mathbb{R}^{d_{\text{model}}}$ are feature-specific parameters. Categorical features are embedded through trainable lookup tables. A learnable [CLS] token $\mathbf{t}_{\text{cls}}$ is prepended to the resulting sequence $[\mathbf{t}_{\text{cls}}, \mathbf{t}_1, \dots, \mathbf{t}_D]$, which is processed by an L -layer Transformer encoder with $d_{\text{model}}=192$, $H=4$ attention heads, and pre-layer normalization (Xiong et al., 2020). The [CLS] output is projected through a two-layer feed-forward network (FFN) to produce the EHR embedding:

$$\mathbf{z}_{\text{EHR}} = \text{FFN}\left(\text{Transformer}^L([\mathbf{t}_{\text{cls}}, \mathbf{t}_1, \dots, \mathbf{t}_D])_{\text{cls}}\right) \in \mathbb{R}^{512}. \quad (1)$$

To capture short-horizon physiological trends beyond the index measurement, 96-hour prior vital sign histories are summarized per channel into seven aggregate statistics (minimum, maximum, mean, standard deviation, first, last, and slope), yielding 42 supplementary features concatenated to the base vector before encoding. Missing values are imputed with training-fold medians to prevent information leakage across data splits. The EHR encoder is trained end-to-end during Stage 1 and frozen thereafter.

Input windows. Two windows govern the inputs to PRISM, and both are defined relative to ED arrival, denoted t_0 . The MDS-ED benchmark extracts its feature vector from the interval $[t_0, t_0 + 90 \text{ min}]$, and this interval also fixes the prediction cutoff. No information recorded after $t_0 + 90 \text{ min}$ enters PRISM through any stream, so the model sees nothing that postdates the point of prediction. The supplementary vital sign statistics described above and the prior ECG acquisitions described in Section 3.4 are drawn instead from the retrospective interval $[t_0 - 96 \text{ h}, t_0]$, which lies wholly before arrival. This design extends the evidence available at the point of prediction without crossing the cutoff. Prior systems evaluated on MDS-ED use only the 90-minute post-arrival window, so the pre-arrival record is signal that those systems leave unused. We measure its contribution directly by retraining PRISM under the benchmark window alone, and Section 4.1 reports the resulting matched-window comparison.

3.3 ECG Waveform Encoding

The index-visit ECG waveform $\mathbf{X}_{\text{ecg}} \in \mathbb{R}^{12 \times T}$ is encoded by D-BETA (Hung et al., 2025), a waveform-language foundation model pre-trained through two complementary objectives: masked modality modeling, which reconstructs masked segments of both the waveform and paired clinical text, and a contrastive objective with a nearest-neighbor negative sampling strategy that aligns waveform and text representations in a shared embedding space. The encoder produces a fixed-dimensional waveform embedding $\mathbf{z}_{\text{ECG}} \in \mathbb{R}^{d_{\text{ecg}}}$. D-BETA is selected as default ECG encoder for its joint waveform-text pre-training, which organizes the embedding space by clinical co-occurrence. This geometric property benefits both patient-level encoding and population memory bank retrieval (see Section 3.5). The encoder remains frozen throughout both training stages, with embeddings pre-computed offline.

3.4 Temporal History Aggregation

Serial ECG changes carry prognostic information beyond any single acquisition (Lehmacher et al., 2020). To incorporate this longitudinal context, we introduce a temporal history aggregator that weights prior-visit

ECG embeddings by their recency relative to the index encounter (Figure 1, right panel). For a patient with N prior acquisitions obtained δ_i hours before the index visit, the aggregator computes normalized exponential recency-decay weights:

$$w_i = \frac{\exp(-|\lambda| \delta_i / \tau)}{\sum_{j=1}^N \exp(-|\lambda| \delta_j / \tau)}, \quad (2)$$

where $\lambda \in \mathbb{R}$ is a single learnable scalar initialized to 1.0 and $\tau = 96$ hours is a fixed normalization constant corresponding to the history window. The absolute value $|\lambda|$ enforces non-negative decay while permitting unconstrained gradient flow. The pooled history representation is then:

$$\mathbf{z}_{\text{pool}} = \sum_{i=1}^N w_i \mathbf{z}_{\text{ECG}}^{(i)} \in \mathbb{R}^{d_{\text{ecg}}}. \quad (3)$$

This weighted average is transformed by an adapter network comprising a linear projection, GELU activation, dropout ($p=0.1$), a second linear projection to 512 dimensions, and layer normalization:

$$\mathbf{z}_{\text{hist}} = \text{LayerNorm}(\mathbf{W}_2 \text{Dropout}(\text{GELU}(\mathbf{W}_1 \mathbf{z}_{\text{pool}} + \mathbf{b}_1)) + \mathbf{b}_2) \in \mathbb{R}^{512}. \quad (4)$$

For patients without prior acquisitions ($N=0$), a separate learnable no-history embedding initialized from $\mathcal{N}(0, 0.02^2)$ replaces $\mathbf{z}_{\text{hist}}$, allowing the model to distinguish the absence of history from a genuine temporal representation. The entire aggregation module, including λ , is trained during Stage 1 and frozen in Stage 2. Training moves $|\lambda|$ from its initialization of 1.0 to 0.9891, which corresponds to a half-life of 67.3 hours over the 96-hour window. The resulting weighting is graded rather than saturating. Two acquisitions separated by one hour receive weights in the ratio 1.01:1, at twelve hours 1.13:1, at two days 1.64:1, and across the full window 2.69:1. The aggregator therefore behaves close to a uniform average for acquisitions taken minutes apart and discriminates increasingly as the separation grows to hours and days, reflecting the recency prior when acquisition times are irregular. The scope of the mechanism is set by acquisition availability. For the 89.3% of encounters with no prior acquisition the learned no-history embedding is used, and for the 8.8% with exactly one the normalized weight is 1 irrespective of λ . The pooling rule is therefore operative for the 1.9% of encounters carrying two or more acquisitions, and Section 4.6 evaluates it on exactly that population. Appendix G reports the availability distribution and the derived decay quantities.

3.5 Population Memory Bank

To provide population-level retrieval context, PRISM constructs a dual-modality memory bank using FAISS (Douze et al., 2024) over per-encounter embeddings from the training cohort. After Stage 1 training, the learned EHR and ECG encoders produce fixed-length representations for every training encounter. These embeddings are ℓ_2 -normalized and stored in two independent nearest-neighbor indices: an EHR index operating over $\mathbf{z}_{\text{EHR}} \in \mathbb{R}^{512}$ for EHR-derived similarity, and an ECG index operating over $\mathbf{z}_{\text{ECG}} \in \mathbb{R}^{d_{\text{ecg}}}$ for ECG-derived similarity. At inference, the query patient’s EHR and ECG embeddings are used to retrieve the k nearest neighbors independently from each index using cosine similarity. The top- k retrieved embeddings from each modality are aggregated via mean pooling to produce context vectors:

$$\mathbf{c}_{\text{EHR}} = \frac{1}{k} \sum_{i=1}^k \mathbf{z}_{\text{EHR}}^{(\text{nn}_i)} \in \mathbb{R}^{512}, \quad \mathbf{c}_{\text{ECG}} = \frac{1}{k} \sum_{i=1}^k \mathbf{z}_{\text{ECG}}^{(\text{nn}_i)} \in \mathbb{R}^{d_{\text{ecg}}}, \quad (5)$$

where nn_i denotes the i -th nearest neighbor. The neighborhood size $k=3$ is selected via held-out sensitivity analysis over $k \in \{1, 3, 5\}$ and held fixed across all task groups. To eliminate retrieval latency during training and inference, neighbor assignments are materialized as pre-computed row-aligned lookup tables after Stage 1, reducing each retrieval operation to a single array index. Appendix C provides a complete configuration of the retrieval index and its key details.

Retrieved content. Equation 5 specifies the retrieved context completely. The model receives two vectors per encounter, $\mathbf{c}_{\text{EHR}}$ and $\mathbf{c}_{\text{ECG}}$, and each is the unweighted mean of $k=3$ neighbor embeddings drawn from its own modality-specific index. Retrieval therefore operates entirely in representation space. Stage 1 training

on the prediction objective shapes the geometry of these embeddings, so outcome-related structure may be reflected indirectly in the pooled vectors; however, no label, code, feature value, identifier or descriptive text from a neighbor enters the prompt. The two indices are queried independently and pass through separate projectors into separate soft-token positions, so the decoder receives EHR-derived and ECG-derived population context as distinct inputs.

Same-patient retrieval. The MDS-ED partitions are disjoint at the level of the patient rather than the encounter, and the memory bank is built only from training-fold encounters. A validation or test query therefore cannot return an encounter belonging to the same patient, because no encounter of that patient is present in the index, and the same-patient retrieval rate at evaluation is exactly zero. Within the training fold the situation is different in principle, since a patient with several encounters may retrieve one of their own, and we quantified this directly over all training and validation queries. At the deployed neighborhood size of $k=3$, 29 queries (0.0237%) retrieve an encounter belonging to the same subject, against roughly 0.01% expected under random retrieval. This remains a very small fraction despite 76,875 rows in the combined training and validation cohort (62.7% of 122,641 encounters) belonging to patients with multiple encounters; the cause of this small enrichment is not established. A related qualification concerns the exclusion mechanism. Self-matches are removed by dropping the rank-0 neighbor, which assumes the query row always ranks first against itself. In 0.0073% of queries it does not, because a near-duplicate embedding displaces it. Filtering on encounter identifier would remove this residual entirely. Accordingly, the implemented procedure is described as rank-0 removal rather than guaranteed self-exclusion. Because evaluation is patient-disjoint, the same-subject retrieval rate is 0 at evaluation; the small training-fold rate makes same-patient lookup unlikely to explain the observed test benefit. The dual-index design keeps EHR- and ECG-derived retrieval context separate. As the ablation study in Section 4.6 shows, neither index alone materially improves over no retrieval, whereas supplying both together does. These controls establish an interaction in the deployed PRISM configuration but do not identify its mechanism; an explicit joint-similarity retrieval comparison remains future work.

3.6 Multimodal Fusion and Language Model Adaptation

Soft-token projection. Each modality embedding is mapped to a sequence of n_m soft tokens in the language model’s embedding space through a dedicated two-layer projector:

$$\mathbf{S}_m = \text{Reshape}_{n_m \times d_{\text{llm}}} \left(\mathbf{W}_m^{(2)} \text{GELU} \left(\mathbf{W}_m^{(1)} \text{LayerNorm}(\mathbf{z}_m) + \mathbf{b}_m^{(1)} \right) + \mathbf{b}_m^{(2)} \right), \quad (6)$$

where $\mathbf{W}_m^{(1)} \in \mathbb{R}^{1024 \times d_{\text{in}}}$, $\mathbf{W}_m^{(2)} \in \mathbb{R}^{n_m d_{\text{llm}} \times 1024}$, $n_m=8$ for all five modality streams, and $d_{\text{llm}}=2560$ for MedGemma-4B. The five modality streams are: current ECG ($\mathbf{z}_{\text{ECG}}$), structured EHR ($\mathbf{z}_{\text{EHR}}$), longitudinal ECG history ($\mathbf{z}_{\text{hist}}$), retrieved EHR context ($\mathbf{c}_{\text{EHR}}$), and retrieved ECG context ($\mathbf{c}_{\text{ECG}}$). Appendix B tabulates complete per-modality input and output dimensions for all five projectors.

Placeholder substitution. The tokenized clinical prompt $\mathbf{p}$ contains four or five designated placeholder tokens that serve as positional anchors within the text sequence, the fifth present only when the encounter has at least one prior ECG acquisition. At forward time, each placeholder is replaced in-place by its corresponding $n_m \times d_{\text{llm}}$ soft-token matrix $\mathbf{S}_m$, while all remaining positions are embedded through the standard LLM vocabulary layer. The resulting sequence interleaves 32 or 40 modality tokens (4×8 or 5×8 , according to whether prior ECG history is available) with a variable number of text tokens into a unified embedding tensor $\mathbf{E} \in \mathbb{R}^{S \times d_{\text{llm}}}$ consumed directly by the decoder. This substitution-based alignment enables cross-modal reasoning within the LLM’s self-attention layers without architectural modification.

Language model backbone and adaptation. We use MedGemma-4B (Sellersgren et al., 2025), a decoder-based four-billion-parameter language model pre-trained on de-identified biomedical corpora. During Stage 1, all LLM parameters are frozen and only the encoders and projectors are optimized. During Stage 2, we apply Low-Rank Adaptation (LoRA) (Hu et al., 2022) to the query, key, value, output, gate, up-projection, and down-projection matrices of every self-attention layer with rank $r=16$, scaling factor $\alpha_{\text{LoRA}}=32$, and dropout 0.05.

3.7 Training and Inference

PRISM is trained in two stages, each with a distinct set of trainable parameters and a shared composite objective.

Composite objective. Both stages minimize a weighted combination of causal language modeling loss on the answer tokens and a binary classification auxiliary loss:

$$\mathcal{L} = \mathcal{L}_{\text{CLM}} + \alpha \mathcal{L}_{\text{BCE}}, \quad (7)$$

where $\mathcal{L}_{\text{CLM}}$ is the standard next-token cross-entropy restricted to the answer span, $\mathcal{L}_{\text{BCE}}$ is binary cross-entropy applied through a linear classification head over the mean-pooled decoder hidden states at all prompt positions, and $\alpha=0.1$ throughout.

Stage 1: encoder and projector training. The EHR FT-Transformer, the temporal history aggregator (including λ), and the three primary projectors (ECG, EHR, history) are jointly optimized while the D-BETA encoder and MedGemma-4B backbone remain frozen. Training uses AdamW optimizer with learning rate 2×10^{-4} , cosine decay with 100 warm-up steps, gradient clipping at $\|\mathbf{g}\|_2 \leq 1.0$, effective batch size 32 per device (micro-batch 4, gradient accumulation over 8 steps), and BF16 mixed precision on two NVIDIA H100 80 GB GPUs. Upon completion, the population memory bank indices are constructed from the trained encoder outputs.

Stage 2: retrieval-augmented language model adaptation. All Stage 1 parameters are frozen. The LoRA adapters, the two retrieval projectors (initialized from scratch), and the classification head (loaded from the Stage 1 checkpoint) are jointly optimized. The learning rate is reduced to 2×10^{-5} ; the micro-batch size is reduced to two to accommodate the expanded five-modality token sequence within the 2,048-token budget. Checkpoints are selected by minimum validation loss on the held-out split. Full hyperparameter specifications are provided in Appendix A.

Inference pathways. At inference, PRISM supports two complementary output pathways evaluated on every task. The *generative pathway* applies a softmax to the Yes/No answer-token logits at the final prompt position and uses the Yes probability as the prediction score, reported as $\text{PRISM}_{\text{gen}}$. The *discriminative pathway* applies the linear head to the mean-pooled decoder hidden states, producing a continuous probability estimate $\hat{y} = P(y=1) \in [0, 1]$ without autoregressive decoding, reported as $\text{PRISM}_{\text{disc}}$. Both pathways are reported for the primary benchmark; the external and temporal comparisons report the generative pathway.

4 Results

4.1 Overall Benchmark Performance

The primary benchmark experiments use the standardized MDS-ED splits; external evaluation is reported in Section 4.2, with supplementary temporal validation in Appendix N. Following Stage 1 and Stage 2 optimization on 4.5 million training samples, the benchmark test partition contains 6,454 encounters. The sampled evaluation retains 262,158 encounter–task records after removing records from 83 encounters with unavailable ECG waveforms; exact class balance is therefore not guaranteed after exclusion. Table 1 reports AUROC across all four MDS-ED task groups for PRISM and 15 baseline systems spanning five model families. PRISM attains the strongest reported figures on every task group: **86.24%** on Diagnosis (1,428 ICD-10 discharge tasks), **93.12%** on Deterioration (6 tasks), **91.79%** on ICU admission (2 tasks), and **92.82%** on Mortality (7 tasks), with the generative and discriminative pathways achieving the best score on complementary task groups. The supervised multimodal systems have higher diagnosis AUROC than the listed single-modality ECG models and zero-shot language models (Table 1). Among the four task groups, Deterioration, ICU admission, and Mortality predict future adverse events from data available within the first 90 minutes after ED arrival, whereas Diagnosis tasks predict discharge ICD-10 labels assigned at the end of the hospital stay, a temporally forward-looking classification problem grounded in a different clinical

Table 1: **Benchmark results across all 1,443 MDS-ED tasks.** AUROC (%) is macro-averaged within each task group: Diagnosis (1,428 ICD-10 discharge tasks), Deterioration (6 tasks), ICU admission (2 tasks), and Mortality (7 tasks). PRISM has the highest reported values among the listed methods in each task group. Its full-input configuration additionally uses 96-hour pre-arrival history; the matched-window comparison is reported in the text. **Bold with underline**: best. [†]Zero-shot, text-only input.

Method	Tasks · AUROC (%)			
	Diagnosis	Deterioration	ICU	Mortality
ECG Foundation Models · <i>waveform input</i>				
HeartBERT (Tahery et al., 2025)	56.35	58.24	57.28	61.11
ECGFM-KED (Tian et al., 2024)	58.75	64.76	61.32	65.76
ESI (Yu et al., 2024)	60.98	63.34	62.52	66.28
ECGFounder (Li et al., 2025)	61.52	69.35	64.13	69.42
Image-Based Models · <i>rendered ECG input</i>				
GEM (Lan et al., 2025)	61.52	55.80	60.40	65.70
PULSE (Liu et al., 2024)	58.47	61.10	64.70	68.70
TimeMaster (Zhang et al., 2025)	57.91	60.80	63.80	57.30
ECG-Instruct (M & M, 2025)	47.33	46.00	44.40	43.70
Signal-Based QA Models · <i>ECG waveform input</i>				
ECGQA-FSL (Tang et al., 2025)	50.56	50.46	51.06	52.15
Q-HEART (Pham et al., 2025)	50.15	55.42	54.89	56.32
ECG-Chat (Zhao et al., 2025)	49.70	57.54	56.40	55.78
Frontier LLMs [†] · <i>text-only input</i>				
GPT-5 [†] (OpenAI, 2025)	55.68	78.12	70.04	62.27
Gemini-2.5 Pro [†] (Comanici et al., 2025)	57.82	78.72	70.70	72.38
Multimodal Baselines · <i>waveform + EHR</i>				
MDS-ED (Lopez Alcaraz et al., 2025)	82.56	90.70	90.63	91.68
UniPACT (Tang et al., 2026)	83.98	91.17	90.50	91.82
PRISM (Ours) · <i>waveform + EHR + history + retrieval</i>				
PRISM_{gen}	<u>86.24</u>	<u>92.90</u>	<u>91.79</u>	<u>92.82</u>
PRISM_{disc}	<u>86.06</u>	<u>93.12</u>	<u>91.55</u>	<u>92.79</u>

signal. Table 1 and Table 11 report the single-run evaluation. The three-seed evaluation used for the control experiments yields group means of 0.8619, 0.9291, 0.9180, and 0.9279, corresponding to a macro-of-macro AUROC of 0.9092, compared with 0.9094 in the single-run evaluation. This 0.02-point difference is within the 0.32-point standard deviation across the three full-model runs. Ablation effects are computed relative to the corresponding full-model reference from the corresponding evaluation.

The two prior multimodal systems, MDS-ED (Lopez Alcaraz et al., 2025) and UniPACT (Tang et al., 2026), fuse ECG representations with clinical context and reach 88.89% and 89.37% overall. Relative to UniPACT, the highest PRISM scores across the two pathways are higher by 2.26 points on Diagnosis, 1.95 on Deterioration, 1.29 on ICU admission, and 1.00 on Mortality.

PRISM uses the 96-hour pre-arrival window described in Section 3.2, whereas the reported baselines use only the benchmark window. Removing this additional history and retraining reduces PRISM from 0.9092 to 0.8895, a difference of 1.97 AUROC points. Under the benchmark window alone, performance is closely spaced across PRISM (0.8895), UniPACT (0.8937), and MDS-ED (0.8889), indicating that the additional pre-arrival history contributes materially to PRISM’s full-input benchmark result. Section 4.5 provides a complementary matched-input comparison against compact non-LLM fusion models receiving the same evidence streams and frozen upstream encoders. PRISM outperforms the tested compact configurations; the comparison and its methodological scope are detailed in Section 4.5.

Table 2: **External evaluation on MC-MED.** All entries are AUROC; PRISM uses the generative pathway. Overall AUROC is the unweighted mean of the four task-group means. Brackets show reported 95% patient-bootstrap confidence intervals. The first two rows use no target-site parameter training. The final two rows perturb retrieval at inference using the fine-tuned checkpoint and identical evaluation records. Bold identifies the highest point estimate among the supervised rows, not statistical significance.

Model / condition	Diagnosis	Deterioration	ICU	Mortality	Overall AUROC
<i>Source-only transfer</i>					
PRISM, original bank	0.7086	0.7112	0.8188	0.8977	0.7841 [0.751, 0.814]
<i>Local-bank adaptation (no parameter updates)</i>					
PRISM, MC-MED bank	0.7080	0.7133	0.8193	0.9025	0.7858 [0.754, 0.815]
<i>Supervised local training / adaptation</i>					
Gradient boosting	0.7012	0.7509	0.8924	0.9465	0.8227 [0.796, 0.851]
Fusion MLP	0.6184	0.6227	0.7188	0.8150	0.6937 [0.658, 0.729]
Fusion MLP + retrieval	0.5847	0.6471	0.6987	0.7469	0.6694 [0.635, 0.703]
Fine-tuned PRISM	0.8445	0.7690	0.8830	0.9479	0.8611 [0.834, 0.887]
<i>Inference-time controls of fine-tuned PRISM</i>					
Retrieval zeroed	0.7920	0.7389	0.8681	0.9215	0.8301 [0.802, 0.857]
Random retrieval	0.7919	0.7310	0.8657	0.9236	0.8280 [0.799, 0.856]

We estimated 95% confidence intervals via empirical bootstrap (1,000 resamples per task, macro-averaged across tasks within each group). The resulting intervals are 86.24 [85.92, 86.45] for Diagnosis, 92.90 [89.16, 95.82] for Deterioration, 91.79 [90.92, 92.65] for ICU, and 92.82 [91.98, 93.50] for mortality. A complete breakdown of per-task performance for Deterioration, ICU, and Mortality, with confidence intervals, is provided in Appendix D.

4.2 External Evaluation on MC-MED

We further evaluate PRISM on MC-MED (Kansal et al., 2025) from a separate institution. Table 2 distinguishes source-only transfer using the original checkpoint and memory bank, local-bank adaptation without parameter updates, and supervised local adaptation by fine-tuning PRISM. The locally trained baselines provide the primary comparison for fine-tuned PRISM; they receive target-site supervision unavailable in the source-only regime. All models are evaluated on the same eligible records and tasks within this setting.

MC-MED comprises Stanford emergency-department visits from 2020–2022. Its single measured ECG lead is repeated across the 12 encoder input channels, so this evaluation also involves a change in waveform acquisition rather than an identical diagnostic 12-lead input. Its chronological split contains 31,239 training, 4,579 validation, and 4,303 test visits. Evaluation uses 13,254 class-balanced encounter–task records from 3,588 patients across 306 eligible tasks (297 diagnosis, four deterioration, one ICU, and four mortality tasks). PRISM results in Table 2 use the generative pathway. Appendix M provides cohort counts, endpoint coverage, and reporting conventions.

Replacing the original bank with an MC-MED bank changes overall AUROC from 0.7841 to 0.7858, an increase of 0.17 points. Supervised fine-tuning reaches 0.8611, an increase of 7.70 points over source-only transfer. Fine-tuned PRISM exceeds locally trained gradient boosting by 3.84 points overall, with the largest task-group advantage in diagnosis. Gradient boosting retains the higher ICU point estimate, and mortality performance is close. These results support supervised adaptation to the external setting rather than a claim that bank replacement alone resolves institutional shift. Inference-time retrieval controls provide a complementary comparison within the fine-tuned model. Zeroing retrieval lowers overall AUROC from 0.8611 to 0.8301, while random neighbors yield 0.8280, corresponding to decreases of 3.10 and 3.31 AUROC

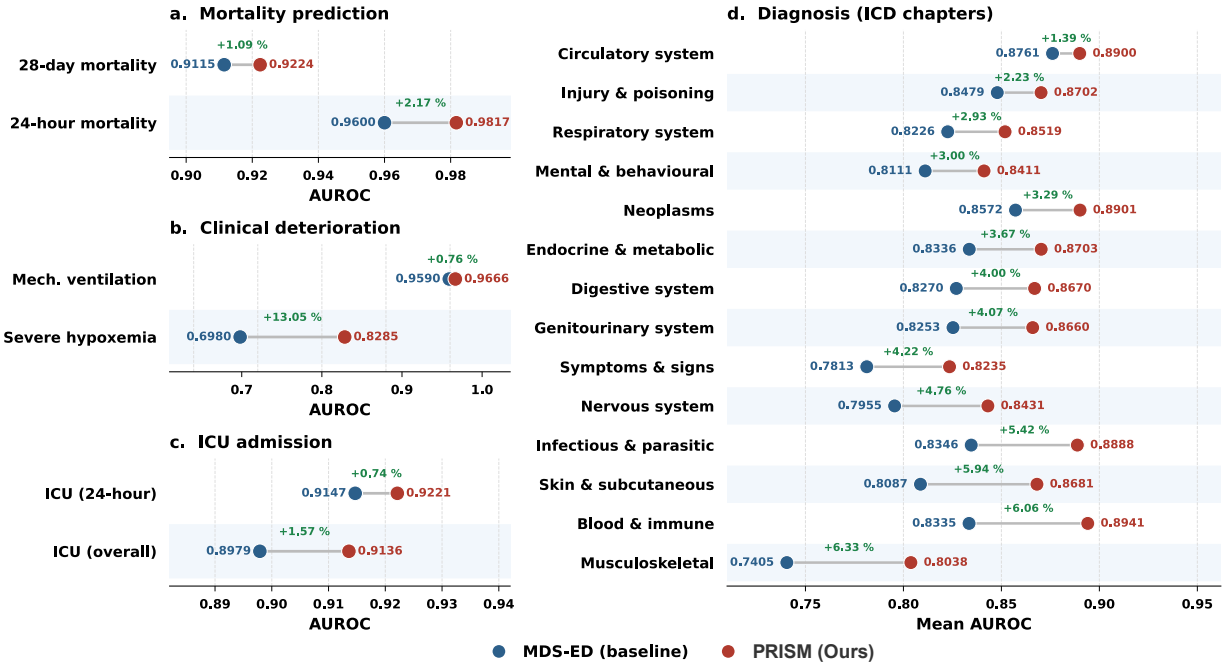


Figure 2: **PRISM performance vs. MDS-ED baseline across clinical prediction categories.** Comparison of AUROC for (a) mortality at multiple horizons, (b) clinical deterioration endpoints, (c) ICU admission, and (d) ICD-chapter-level diagnosis groups. Red and blue markers denote PRISM and MDS-ED, respectively. Green annotations indicate relative AUROC improvement. Gains are largest on laboratory-dependent outcomes (e.g., blood and immune disorders, musculoskeletal conditions) and on deterioration endpoints where metabolic measurements are clinically relevant.

points. Genuine retrieval has higher point estimates in all four task groups. The two controls differ by only 0.21 points overall; this small difference does not establish their equivalence. The pattern is consistent with useful information in the selected neighbor representations, without identifying how the model uses that information. Group-specific intervals are reported in Appendix M, Table 22.

4.3 Task-Level Analysis

Task-level analysis reinforces the interpretation that PRISM’s gains are concentrated where structured clinical data carries the most discriminative weight. Across the full benchmark we count tasks whose lower 95% confidence bound exceeds an AUROC of 0.80, following the criterion used by UniPACT so that the comparison is like-for-like. PRISM qualifies on 1,145 of 1,443 tasks (79.3%), against 883 for UniPACT (61.2%) and 623 for MDS-ED (43.2%). This is a gain of 262 tasks over UniPACT and 522 over MDS-ED. At the level of ICD chapters (Figure 2d), improvements are consistent across all clinical domains but graded by the degree to which each domain depends on laboratory context. Circulatory system diagnoses, where ECG waveforms already carry substantial discriminative information, show a gain of 1.39%. In contrast, musculoskeletal disorders (+6.33%) and blood and immune system conditions (+6.06%) show the largest improvements, consistent with these categories drawing strongly on laboratory measurements for which the numerical EHR pathway may be particularly useful. Within the Deterioration group, the most striking individual result is on severe hypoxemia, where PRISM achieves an AUROC of 0.8285 compared to 0.6980 for MDS-ED (+13.05%). Arterial oxygenation and respiratory measurements are clinically relevant to this endpoint and may help explain the larger observed gap. Appendix F provides a comprehensive view of how these gains distribute across the 1,428 diagnosis tasks and ICD-10 chapters.

4.4 Population Memory Bank Analysis

Figure 3 presents a representative case study examining the clinical coherence of the population memory bank. For a test patient assigned ICD-10 label Z79 (long-term drug therapy) with a positive outcome, we

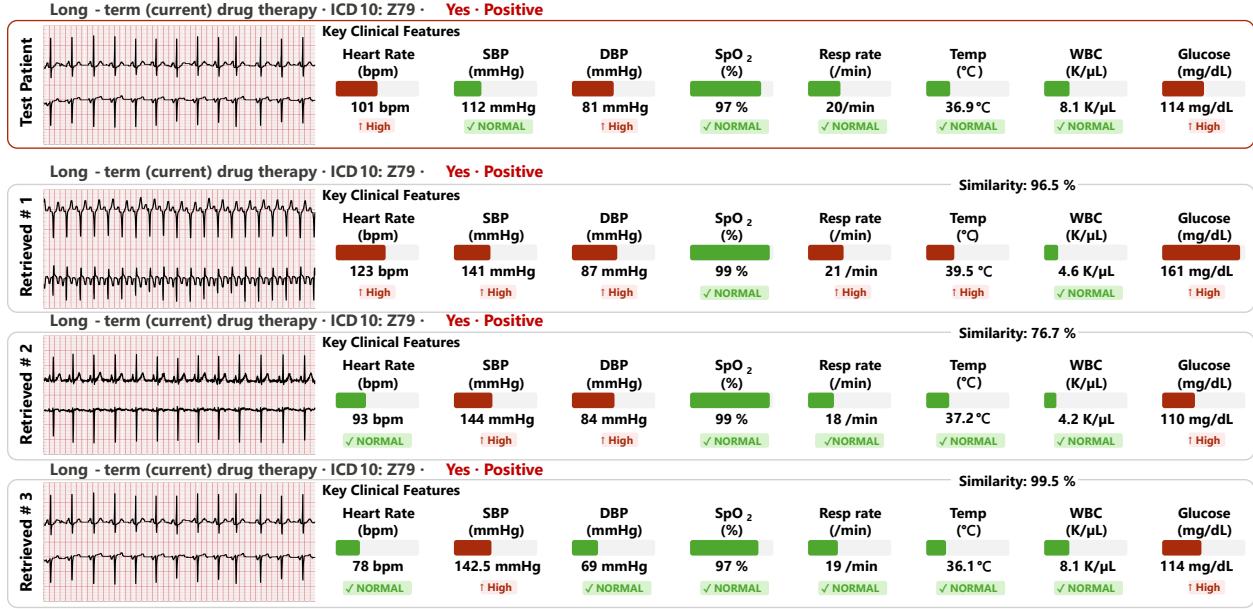


Figure 3: **Qualitative analysis of the population memory bank.** A test encounter with ICD-10 label Z79 (long-term drug therapy, positive outcome) and three training-fold encounters selected from the dual-modality memory bank for illustration. The EHR and ECG indices are queried independently; the displayed encounters are returned by both indices. The model itself receives the two mean-pooled retrieval vectors defined in Equation 5, rather than the individual encounters shown here. All displayed encounters share the same discharge label and outcome. Similarity values are reported within the index from which they are obtained and are not directly comparable across modalities. Shared abnormalities, including elevated glucose and heart rate, illustrate local clinical coherence in the learned embedding space.

inspect the three training encounters whose EHR and ECG embeddings are nearest to the query in each index, with the cosine similarities shown measured within the index that returned each neighbor, since a neighbor has one similarity to the query per index and the two are not commensurable. All three retrieved neighbors share both the discharge label and the positive outcome, which illustrates local structure in the learned embedding space. The decoder conditions on the pooled vectors of Equation 5 rather than on these individual records, so the case is descriptive rather than evidence of cohort-level outcome concordance. The controls in Section 4.6 address the prediction mechanism directly: genuine retrieved content improves the aggregate result relative to random and zeroed retrieval. Inspection of key clinical features reveals plausible physiological correspondences: the test patient and all retrieved neighbors share elevated glucose and elevated heart rate as flagged abnormalities, consistent with the metabolic context associated with long-term pharmacological management in this diagnostic category. This selected case illustrates local coherence in the embedding space and does not support cohort-level inference. Because the index stores representations of real encounters, the underlying neighbors can be enumerated and reviewed offline, providing retrieval provenance without placing individual retrieved records, labels, or outcomes in the model prompt. Extended retrieval examples across four clinical phenotypes, namely metabolic disorder, chronic vascular, endocrine management and vascular risk, are provided in Appendix L, where selected $k=50$ neighborhoods illustrate the index geometry. The larger neighborhood is used only to make prevalence estimates readable, and no cohort-level inference is drawn from these examples.

4.5 Comparison with Compact Non-LLM Models

To complement the within-PRISM ablations, we train four compact non-LLM controls using the same upstream encoders where applicable, together with a direct k NN outcome-prevalence baseline. The fusion+retrieval control receives the same five input streams as PRISM, with the test fold and evaluation protocol held fixed. Table 4 reports the comparison.

The final compact control receives all five streams, including both retrieval contexts, and is therefore matched to PRISM on input information. PRISM exceeds this 5.9M-parameter control by 14.40 AUROC points.

Table 3: **Architectural and training ablations.** AUROC is reported as macro-of-macro over the four task groups under the generative pathway. Rows marked $\dagger$ use the three-seed full-model reference (0.9092 ± 0.0032); unmarked rows use the single-run reference (0.9094). Individual ablations are single training runs unless stated otherwise. Δ is calculated against the corresponding full-model reference.

Ablation / configuration	AUROC	Δ
Encoder ablations		
Text-serialized EHR (no tokenization) $\dagger$	0.8726	−3.66
w/o EHR encoder	0.8874	−2.20
w/o current-visit ECG $\dagger$	0.8931	−1.61
w/o longitudinal ECG history $\dagger$	0.8970	−1.22
Memory bank ablations, three-seed reference		
w/o retrieval context $\dagger$	0.8993	−0.99
Retrieval with random neighbors $\dagger$	0.8995	−0.97
Memory bank ablations, single-run reference		
w/o retrieval context	0.8904	−1.90
EHR index only	0.8910	−1.84
ECG index only	0.8906	−1.88
Neighborhood size		
$k = 1$	0.8911	−1.83
$k = 5$	0.8932	−1.62
Training ablations		
Frozen retrieval projectors	0.8704	−3.90
LoRA off (backbone frozen)	0.8219	−8.75
Full PRISM, three-seed reference$\dagger$	0.9092	—
Full PRISM, single-run reference	0.9094	—

Table 4: **Architectures with and without a language model.** The fusion+retrieval control is matched to PRISM on all five input streams; the remaining compact controls use the subsets indicated in their row labels. Frozen upstream encoders are shared where applicable, and all models use the same test fold and evaluation protocol. AUROC is reported as mean $\pm$ standard deviation over five seeds for the compact models and over three seeds for PRISM. Expected calibration error uses ten equal-count bins weighted by bin size. Average precision is reported at true cohort prevalence. Parameter counts are totals including frozen components.

Model	LLM	AUROC	AUPRC	ECE	Params
k NN outcome prevalence ($k=50$)	No	0.5262	0.1235	—	0
FT-Transformer, structured EHR only	No	0.6870 ± 0.0058	0.1864	0.4362	3.1M
ECG and longitudinal history only	No	0.7527 ± 0.0031	0.2378	0.4308	4.1M
Fusion, EHR + ECG + history	No	0.7759 ± 0.0051	0.2727	0.4261	4.6M
Fusion + retrieval (all five streams)	No	0.7652 ± 0.0043	0.2500	0.4043	5.9M
PRISM	Yes	0.9092 ± 0.0032	0.5504	0.1134	4B

Against the strongest compact configuration, the 4.6M-parameter three-stream model without retrieval, the margin is 13.33 points; average precision increases from 0.2727 to 0.5504 and expected calibration error decreases from 0.4261 to 0.1134. These comparisons show that the compact fusion architectures tested do not exploit the available inputs as effectively as PRISM, but they do not isolate the language model itself because model class, pretrained capacity, parameter count, and optimization also differ. Retrieval changes the compact fusion model from 0.7759 to 0.7652, and a direct k NN outcome-prevalence predictor reaches 0.5262. In these controls, retrieval does not improve the compact fusion architecture, and the direct k NN rule performs near chance; both observations are specific to the tested configurations.

Table 5: **Contribution of each evidence stream.** One evidence stream is removed or replaced per row, with the remaining model components held fixed. Δ is the change in AUROC relative to full PRISM, in points. Random-neighbor and zeroed-token controls retain the retrieval architecture while varying the content supplied through it.

Configuration	AUROC	AUPRC	ECE	Δ
Structured EHR as serialized text	0.8726	0.4632	0.1586	−3.66
w/o 96-hour pre-arrival window	0.8895	0.5358	0.1290	−1.97
w/o current-visit ECG	0.8931	0.5031	0.1266	−1.61
w/o longitudinal ECG history	0.8970	0.5418	0.1260	−1.22
w/o retrieval (tokens zeroed)	0.8993	0.5450	0.1239	−0.99
Retrieval with random neighbors	0.8995	0.5470	0.1224	−0.97
PRISM (± 0.0032, three seeds)	0.9092	0.5504	0.1134	–

The compact controls also show that both structured EHR and ECG history carry useful signal: the EHR-only model reaches 0.6870, the ECG+history model 0.7527, and their compact fusion 0.7759. PRISM uses a shared-attention soft-token interface over five heterogeneous streams while answering 1,443 task questions without task-specific heads, and Table 4 therefore compares complete model configurations rather than isolating that design choice. The trade-off is substantial: PRISM has roughly three orders of magnitude more parameters than the compact fusion controls. Parameter counts are reported for all configurations as a model-size proxy. Cross-architecture wall-clock time is not used as a standardized compute metric because hardware utilization, batching, and implementation differ. For the directly comparable text-serialization experiment on identical hardware, inference increases from roughly four hours to twenty-four (Section 4.6).

4.6 Ablation Study

Table 5 reports the contribution of individual evidence streams, and Table 3 reports the remaining architectural and training ablations. Retraining PRISM three times gives a standard deviation of 0.0032 AUROC. Because each ablation is a single training run, this standard deviation is used only as a scale reference for the observed differences rather than as a formal standard error or significance test.

Replacing the numerical EHR pathway with full text serialization is the largest tested substitution at 3.66 points. For retrieval, zeroed tokens give 0.8993 and random neighbors 0.8995, a difference of 0.02 points, whereas genuine retrieval reaches 0.9092. Because genuine and random retrieval use the same projectors and token positions, added retrieval capacity alone does not explain the true-versus-random gap. The 0.02-point random-versus-zeroed difference is negligible relative to full-model seed variability and is not interpreted as an effect of the token positions themselves.

A paired comparison across all 1,443 tasks gives a more granular picture. PRISM exceeds the random-neighbor control on 701 tasks and falls below it on 633, with 109 ties, so it wins 52.5% of the 1,334 non-tied comparisons (two-sided exact binomial $p = 0.067$). The macro-average moves by 0.97 points while the median task moves very little, which locates the benefit of retrieved content in a subset of tasks rather than spreading it evenly across the nosology. The aggregate improvement is therefore concentrated in a subset of tasks.

Table 3 presents the remaining component removal experiments.

Structured EHR encoding drives diagnostic performance. Removing the FT-Transformer costs 2.20 AUROC points, with the largest loss on the diagnosis nosology and smaller changes on deterioration and ICU admission. This pattern is consistent with many diagnosis tasks drawing strongly on laboratory measurements such as troponin, creatinine, and glucose. We therefore interpret the structured record as an important complementary source of discriminative information in PRISM.

Numerical soft-token pathway versus full text serialization. Replacing the FT-Transformer soft-token pathway with full text-serialized EHR features reduces AUROC by 3.66 points, from 0.9092 to 0.8726,

Table 6: **ECG encoder backbone ablation.** All other architectural components and hyperparameters are held fixed; only the waveform encoder varies. Best AUROC (%) across both inference pathways is reported per task group. Values are per-task-group percentages from the original single-run evaluation, on the same scale as Table 1 and not on the macro-of-macro scale used in Tables 4, 5 and 3. Δ denotes the gap between D-BETA and the next-best encoder. **Bold:** best per column.

ECG Encoder	Diagnosis	Deterioration	ICU	Mortality
HeartBERT (Tahery et al., 2025)	83.65	91.07	90.79	91.61
ECGFM-KED (Tian et al., 2024)	82.71	90.57	90.27	90.90
ECGFounder (Li et al., 2025)	82.54	91.71	90.60	90.73
D-BETA (PRISM) (Hung et al., 2025)	86.24	93.12	91.79	92.82
Δ vs. next-best	+2.59	+1.41	+1.00	+1.21

the largest tested substitution. The standard textual patient summary is retained in both conditions; this ablation changes the additional structured-EHR pathway rather than comparing numerical input against the absence of text. A 2,048-token budget cannot contain all 470 variables once fully serialized without truncation, so the text condition instead receives the required 8,192-token maximum sequence length. Nothing is truncated; at batch size four it processes 32,768 tokens per step, versus 16,384 for PRISM at batch size eight. Thus truncation and context starvation do not explain the observed gap. The comparison does, however, necessarily differ in sequence length, batch size, and optimization conditions, so we interpret it as a practical comparison between full serialization and the compact numerical soft-token pathway rather than a pure causal isolation of numerical representation. The structured record occupies roughly four thousand serialized tokens versus eight projected soft tokens, and inference on identical hardware increases from approximately four hours to twenty-four under full serialization.

ECG modalities contribute dissociable and complementary signal. Removing the current-visit ECG encoder costs 1.61 points and removing longitudinal history costs 1.22, indicating complementary waveform signal within PRISM. Within the longitudinal stream, we compare recency weighting with simpler pooling on the 122 encounters carrying two or more prior acquisitions, the only population on which the rules differ. The learned rule has the highest point estimate at 0.9136, compared with 0.9069 for the best tested fixed decay, 0.9026 for uniform averaging, 0.8998 for the most recent acquisition only, and 0.8903 for the oldest only. The corresponding differences are 0.67, 1.10, 1.38, and 2.33 AUROC points. The converged $|\lambda| = 0.9891$ remains within 1.1% of its initialization, so we interpret the result as evidence for the tested recency-weighted pooling rule rather than for learnability of the decay constant. Because only 1.9% of encounters are affected and these are single-run subgroup comparisons, we treat the differences as descriptive; across the full test fold the six rules span only 0.09 points, below the 0.32-point full-model seed standard deviation. Appendix G reports both strata in full.

Population retrieval benefits from dual-modality indexing. Activating either the EHR or ECG memory bank alone changes little relative to no retrieval in the single-run evaluation, whereas supplying both genuine modalities produces the full retrieval gain. In that evaluation, no retrieval gives 0.8904, EHR-only 0.8910, ECG-only 0.8906, and full PRISM 0.9094. Using the three-seed reference, zeroed retrieval gives 0.8993, random neighbors 0.8995, and genuine retrieval 0.9092. Each comparison uses its corresponding full-model reference. Because genuine and random retrieval use the same retrieval architecture, the true-versus-random difference supports a content-dependent effect. The observed benefit appears only when both genuine modalities are supplied, but these controls do not establish why the interaction occurs. No joint-similarity index or explicit intersection over neighbor sets is computed, and comparison with an explicit joint retrieval geometry remains future work. Same-subject retrieval occurs in 0.0237% of training queries at $k=3$ and is impossible at evaluation because the folds are patient-disjoint, making it unlikely to explain the observed test benefit. Neighborhood size is a descriptive sensitivity result: $k=3$ has the highest point estimate among $\{1, 3, 5\}$, without establishing why the alternatives perform differently.

Training-stage ablations. Freezing the retrieval projectors costs 3.90 AUROC points and disabling LoRA adaptation costs 8.75, the largest single change among the tested ablations. These results show that trainable

Table 7: **Per-group discrimination, average precision and calibration.** AP@eval is measured at the retained evaluation prevalence after ECG exclusions; its chance reference is the retained positive fraction for each task. AP@true rescales to the cohort prevalence of each endpoint. ECE uses ten equal-count bins weighted by bin size.

Group	Tasks	Rows	AUROC	AP@eval	AP@true	Brier	ECE
<i>PRISM_{gen}</i>							
Diagnoses	1,428	253,055	0.8619	0.8532	0.3400	0.1720	0.1558
Deterioration	6	729	0.9291	0.9357	0.5855	0.1186	0.1183
ICU	2	3,476	0.9180	0.9204	0.7265	0.1383	0.0954
Mortality	7	4,898	0.9279	0.9172	0.5495	0.1126	0.0840
Macro-of-macro	1,443		0.9092		0.5504		0.1134
<i>PRISM_{disc}</i>							
Diagnoses	1,428	253,055	0.8599	0.8499	0.3232	0.2015	0.1875
Deterioration	6	729	0.9293	0.9371	0.5750	0.1356	0.1319
ICU	2	3,476	0.9160	0.9077	0.6642	0.1628	0.1367
Mortality	7	4,898	0.9272	0.9153	0.5551	0.1331	0.1132
Macro-of-macro	1,443		0.9081		0.5294		0.1423

retrieval alignment and LoRA adaptation are both important in the tested PRISM configuration; they do not by themselves identify a unique failure mechanism.

ECG encoder selection. Table 6 evaluates three alternative ECG encoder backbones within PRISM with the remaining components held fixed. D-BETA has the highest point estimate on every task group, with margins of +2.59 points on Diagnosis and +1.41 on Deterioration over the next-best encoder. D-BETA’s joint waveform–text pre-training is compatible with the retrieval setting, but this ablation varies the encoder backbone as a whole and does not isolate why D-BETA performs better.

4.7 Control Summary

Random-neighbor and zeroed retrieval differ by only 0.02 AUROC points, below the 0.32-point full-model seed standard deviation, while genuine retrieval is higher than both. Neighborhood size and history pooling also show sensitivity to the tested choices, with $k=3$ and the learned decay rule giving the highest point estimates among their respective alternatives. Because k was selected on held-out performance, λ was learned, and the pooling comparisons affect only a small subgroup, we treat these patterns as descriptive controls rather than causal evidence.

4.8 Additional Metrics and Subgroup Analysis

Many of the 1,443 tasks are substantially imbalanced, so AUROC alone does not fully characterize performance, and we therefore report AUPRC and calibration alongside it. Table 7 reports discrimination, average precision and calibration for each task group under both inference pathways. Average precision appears twice, at the retained evaluation prevalence after ECG exclusions and corrected to true cohort prevalence, since only the latter is clinically interpretable. The two pathways are close throughout, with the generative pathway slightly ahead on aggregate, and calibration is best on the ICU and mortality endpoints and weakest on the diagnosis nosology.

We report subgroup discrimination and calibration in Appendix I, but we do not evaluate calibration under temporal or cross-site distribution shift. AUROC declines across the age bands from 0.9277 in patients under 40 to 0.8684 in those aged 85 and over. Twenty of the twenty-one pairwise subgroup interval comparisons overlap; the exception is the 40–54 versus 85+ age bands, and no sex or self-reported ethnicity comparison separates. Excluding the Hispanic/Latino subgroup, subgroup AUROC is associated with subgroup event rate ($r = -0.78$), as detailed in Appendix I. This descriptive association is consistent with differences in

case mix but does not establish their cause. We therefore treat these analyses as descriptive rather than as evidence of demographic equivalence or disparity.

5 Discussion and Conclusion

Across the evaluations, PRISM’s performance reflects both the evidence available to the model and how that evidence is integrated. Matching the benchmark observation window brings PRISM close to prior multimodal systems, while comparisons with the tested compact fusion models favor PRISM under matched inputs. On MC-MED, supervised fine-tuning contributes a larger improvement than replacing the memory bank alone. Together, these findings motivate evaluating representation, observation window, and local adaptation as distinct design choices.

The retrieval controls further distinguish the presence of context from the selection of useful neighbors. Genuine retrieval exceeds random retrieval by 0.97 AUROC points on MDS-ED, 0.05 points in the temporal evaluation, and 3.31 points after MC-MED fine-tuning. These evaluations differ in cohort, eligible tasks, and training configuration, so the contrasts do not isolate the cause of this variation. Their practical implication is to assess retrieved content against random and zeroed controls in the intended setting, rather than assuming that a memory bank will improve performance.

PRISM extends retrieval augmentation from text corpora to learned representations of training-cohort encounters. In the deployed model, only mean-pooled embeddings enter the language model; retrieved-neighbor labels and individual records do not. The underlying neighbors remain enumerable offline, providing retrieval provenance for retrospective analysis without exposing individual records through the prompt. How such provenance should be presented to end users remains an open deployment question.

Limitations. The evaluation remains retrospective, spans two institutions, and uses a single language-model backbone, MedGemma-4B. Broader multi-site and prospective evaluation is needed to characterize robustness to acquisition, measurement, and population differences. The strongest external result uses supervised local adaptation, and the temporal baseline comparisons are restricted to the tested compact fusion architectures. Subgroup analyses in Appendix I describe performance within the evaluated cohort, while calibration under distribution shift and the clinical use of retrieved evidence remain to be assessed.

Conclusion. PRISM integrates numerical clinical features, ECG waveforms, longitudinal history, and retrieved encounter representations within a multimodal language model. The MDS-ED experiments and supervised MC-MED adaptation support our approach to heterogeneous evidence integration, while the matched-window and retrieval controls identify where additional inputs and selected neighbor representations contribute. These findings motivate evaluation of the complete evidence pipeline when adapting multimodal models to new clinical settings.

Broader Impact Statement

Retrieving over a patient population rather than a curated corpus introduces two classes of risk, and both concern deployment rather than the retrospective evaluation reported here. The first is disclosure. PRISM passes only mean-pooled retrieval embeddings into the language model, and no retrieved-neighbor attribute, identifier, discharge code or outcome is surfaced through the retrieval pathway, which limits direct exposure of individual neighbor records. That property belongs to the current interface rather than to the method, and it would not survive a deployment that displayed individual retrieved cases in order to justify a prediction. Such an interface would expose patient attributes and documented outcomes directly, and it would raise the risk of re-identification through rare combinations of clinical findings, since a neighbor matched closely on an unusual presentation may be effectively unique in the source cohort. Membership inference remains a consideration even where only latent vectors are retained, because an index built over real encounters may reveal whether a given patient contributed to it. A deployment that surfaces retrieved cases would therefore require privacy analysis that this work does not provide. The second risk is interpretive. Similarity in a learned embedding space is not evidence of shared causation. Neighbors who resemble a patient on measured features may differ in treatment history, unrecorded comorbidity, demographic composition or

site-specific practice, and presenting them as support for a prediction may encourage clinicians to reason by analogy from cases that are alike only in the respects the encoder happens to represent. We would regard clinical deployment as premature until subgroup performance has been characterized, until calibration has been assessed under distribution shift rather than on a matched retrospective split, and until the framing of retrieved evidence to end users has been evaluated prospectively.

Acknowledgments

This work was supported by the NWO AiNed Fellowship Grant of A.S., and in part by Google.org and the Google Cloud Research Credits program through the Gemini Academic Program. We also acknowledge the use of the Dutch National Supercomputer Snellius for essential computational tasks.

References

- Amna N Ahmed, Ken Blonde, Daniel Hackam, Alla Iansavichene, and Marko Mrkobrada. Prognostic significance of elevated troponin in non-cardiac hospitalized patients: A systematic review and meta-analysis. *Annals of Medicine*, 46(8):653–663, 2014. doi: 10.3109/07853890.2014.959558.
- Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, and Others. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. URL <https://arxiv.org/abs/2507.06261>.
- Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. 2024.
- Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.
- Manh Pham Hung, Aaqib Saeed, and Dong Ma. Boosting masked ECG-text auto-encoders as discriminative learners. In *Forty-second International Conference on Machine Learning*, 2025.
- Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023. doi: 10.1038/s41597-022-01899-x.
- Aman Kansal, Emma Chen, Tom Jin, Pranav Rajpurkar, and David Kim. Multimodal Clinical Monitoring in the Emergency Department (MC-MED). *PhysioNet*, September 2025. doi: 10.13026/wvyw-g663. URL <https://doi.org/10.13026/wvyw-g663>. Version 1.0.1.
- Ole Kruse, Niels Grunnet, and Charlotte Barfod. Blood lactate as a predictor for in-hospital mortality in patients admitted acutely to hospital: a systematic review. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*, 19:74, 2011. doi: 10.1186/1757-7241-19-74.
- Xiang Lan, Feng Wu, Kai He, Qinghao Zhao, Shenda Hong, and Mengling Feng. GEM: Empowering MLLM for grounded ECG understanding with time series and images. *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.
- Jonas Lehmacher, Johannes Tobias Neumann, Nils Arne Sørensen, Alina Goßling, Paul Michael Haller, Tau Sarra Hartikainen, Peter Clemmensen, Tanja Zeller, Stefan Blankenberg, and Dirk Westermann. Predictive value of serial ECGs in patients with suspected myocardial infarction. *Journal of Clinical Medicine*, 9(7):2303, 2020. doi: 10.3390/jcm9072303.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

-
- Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, volume 36, 2023.
- Jun Li, Aaron D Aguirre, Valderly Moura Junior, Jiarui Jin, Che Liu, Lanhai Zhong, Chenxi Sun, Gari Clifford, M Brandon Westover, and Shenda Hong. An electrocardiogram foundation model built on over 10 million recordings. *NEJM AI*, 2(7):AIoa2401033, 2025.
- Yikuan Li, Shishir Rao, Jose Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Thomas Lukasiewicz, Gholamreza Salimi-Khorshidi, and Kazem Rahimi. BEHRT: Transformer for electronic health records. *Scientific Reports*, 10(7155):1–12, 2020. doi: 10.1038/s41598-020-62922-y.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.
- Ruoqi Liu, Yuelin Bai, Xiang Yue, and Ping Zhang. Teach multimodal LLMs to comprehend electrocardiographic images. *arXiv preprint arXiv:2410.19008*, 2024.
- Juan Miguel Lopez Alcaraz, Hjalmar Bouma, and Nils Strodthoff. Enhancing clinical decision support with physiological waveforms—a multimodal benchmark in emergency care. *Computers in Biology and Medicine*, 192:110196, 2025. doi: 10.1016/j.compbiomed.2025.110196.
- Nandakishor M and Anjali M. High-accuracy ecg image interpretation using parameter-efficient lora fine-tuning with multimodal llama 3.2, 2025. URL <https://arxiv.org/abs/2501.18670>.
- OpenAI. Openai GPT-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.
- Hung Manh Pham, Jialu Tang, Aaqib Saeed, and Dong Ma. Q-HEART: ECG question answering via knowledge-informed multimodal LLMs. In *Proceedings of the European Conference on Artificial Intelligence (ECAI)*, volume 413 of *Frontiers in Artificial Intelligence and Applications*, pp. 4545–4552, 2025. doi: 10.3233/FAIA251356.
- Sushravya Raghunath, Alvaro E Ulloa Cerna, Linyuan Jing, David P vanMaanen, Joshua V Stough, Dustin N Hartzel, Joseph B Leader, H Lester Kirchner, Martin C Stumpe, Ashraf Hafez, Arun Nemani, Tanner Carbonati, Kipp W Johnson, Katelyn Young, Christopher W Good, John M Pfeifer, Aalpen A Patel, Brian P Delisle, Amro Alsaïd, Dominik Beer, Christopher M Haggerty, and Brandon K Fornwalt. Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network. *Nature Medicine*, 26(6):886–891, 2020. doi: 10.1038/s41591-020-0870-z.
- Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-BERT: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digital Medicine*, 4(86):1–13, 2021. doi: 10.1038/s41746-021-00455-y.
- Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atila Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, and others. MedGemma technical report. *arXiv preprint arXiv:2507.05201*, 2025. URL <https://arxiv.org/abs/2507.05201>.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023. doi: 10.1038/s41586-023-06291-2.
- Gary B Smith, David R Prytherch, Paul Meredith, Paul E Schmidt, and Peter I Featherstone. The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. *Resuscitation*, 84(4):465–470, 2013. doi: 10.1016/j.resuscitation.2012.12.016.
- Saadeh Tahery, Fatemeh Hamid Akhlaghi, and Termeh Amirsoleimani. Heartbert: A self-supervised ecg embedding model for efficient and effective medical signal analysis. 2025. URL <https://arxiv.org/abs/2411.11896>.

-
- Jialu Tang, Tong Xia, Yuan Lu, Cecilia Mascolo, and Aaqib Saeed. Electrocardiogram–language model for few-shot question answering with meta learning. In *Proceedings of the Conference on Health, Inference, and Learning (CHIL)*, volume 287 of *Proceedings of Machine Learning Research*, pp. 89–104, 2025.
- Jialu Tang, Tong Xia, Yuan Lu, and Aaqib Saeed. Unipact: A multimodal framework for prognostic question answering on raw ecg and structured ehr. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 22537–22541, 2026. doi: 10.1109/ICASSP55912.2026.11461333.
- Yuanyuan Tian, Zhiyuan Li, Yanrui Jin, Mengxiao Wang, Xiaoyang Wei, Liqun Zhao, Yunqing Liu, Jinlei Liu, and Chengliang Liu. Foundation model of ECG diagnosis: Diagnostics and explanations of any form and rhythm on ECG. *Cell Reports Medicine*, 5:101875, 2024. doi: 10.1016/j.xcrm.2024.101875.
- Dung-Jang Tsai, Yu-Sheng Lou, Chin-Sheng Lin, Wen-Hui Fang, Chia-Cheng Lee, Ching-Liang Ho, Chih-Hung Wang, and Chin Lin. Mortality risk prediction of the electrocardiogram as an informative indicator of cardiovascular diseases. *Digital Health*, 9:20552076231187247, 2023.
- Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. On layer normalization in the transformer architecture. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10524–10533. PMLR, 2020. URL <https://proceedings.mlr.press/v119/xiong20b.html>.
- Han Yu, Peikun Guo, and Akane Sano. ECG semantic integrator (ESI): A foundation ECG model pretrained with LLM-enhanced cardiological text. *arXiv preprint arXiv:2405.19366*, 2024.
- Junru Zhang, Lang Feng, Xu Guo, Yuhan Wu, Yabo Dong, and Duanqing Xu. TimeMaster: Training time-series multimodal LLMs to reason via reinforcement learning. *arXiv preprint arXiv:2506.13705*, 2025.
- Yubao Zhao, Jiaju Kang, Tian Zhang, Puyu Han, and Tong Chen. ECG-Chat: A large ECG-language model for cardiac disease diagnosis. In *IEEE International Conference on Multimedia and Expo (ICME)*, 2025. doi: 10.1109/ICME59968.2025.11209476.
- Yinghao Zhu, Changyu Ren, Zixiang Wang, Xiaochen Zheng, Shiyun Xie, Junlan Feng, Xi Zhu, Zhoujun Li, Liantao Ma, and Chengwei Pan. EMERGE: Enhancing multimodal electronic health records predictive modeling with retrieval-augmented generation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM)*, 2024. doi: 10.1145/3627673.3679582.

Appendix

A. Training Hyperparameters	21
B. Soft-Token Projector Specifications	21
C. Population Memory Bank Configuration	21
D. Per-Task AUROC: Deterioration, ICU, and Mortality	22
E. Diagnosis AUROC by ICD-10 Chapter	22
F. AUROC Distribution Across Diagnosis Tasks	23
G. Longitudinal ECG History: Availability and Aggregation	25
H. Cohort and Dataset Details	25
I. Performance by Patient Subgroup	27
J. Clinical Prompt Design and Representative Examples	28
K. Training Data Construction and Prompt Generation	29
L. Population Memory Bank: Retrieval Across Clinical Phenotypes	31
M. External and Temporal Evaluation Protocols	31
N. Temporal Generalization	33

A Training Hyperparameters

PRISM is trained in two stages with distinct sets of trainable parameters. Stage 1 jointly optimizes the EHR FT-Transformer, the temporal history aggregator (including the learnable decay scalar λ), and the three primary soft-token projectors (ECG, EHR, history), while the MedGemma-4B backbone and D-BETA encoder remain frozen. Stage 2 freezes all Stage 1 parameters and jointly trains the LoRA adapters, the two retrieval projectors (initialized from scratch), and the classification head (loaded from the Stage 1 checkpoint). Table 8 reports all hyperparameters for both stages.

B Soft-Token Projector Specifications

Each modality embedding is mapped to a sequence of eight soft tokens in the language model’s embedding space through a two-layer feed-forward projector with layer normalization, GELU activation, and a final linear expansion (Equation 6). All five projectors share an identical architecture and differ only in input dimensionality and training stage. The three primary projectors (current ECG, EHR, ECG history) are trained during Stage 1. The two retrieval projectors are initialized from scratch and trained during Stage 2 alongside the LoRA adapters. Table 9 summarizes the specifications.

C Population Memory Bank Configuration

The population memory bank comprises two independent nearest-neighbor indices constructed over per-encounter embeddings from the 116,186 training-fold encounters after Stage 1 training. The EHR index operates over 512-dimensional EHR embeddings and the ECG index over d_{ecg} -dimensional ECG embeddings. Both use ℓ_2 -normalized vectors with cosine similarity as the distance metric. To eliminate retrieval latency during training and inference, neighbor assignments for all encounters are materialized as pre-computed row-aligned lookup tables immediately after index construction. At runtime, each retrieval operation reduces to a single array index per modality. The top- k retrieved embeddings from each index are independently mean-pooled to produce context vectors before projection into the language model’s token space. Table 10 summarizes the configuration.

Table 8: Training hyperparameters for PRISM Stage 1 and Stage 2.

Hyperparameter	Stage 1	Stage 2
LLM backbone	MedGemma-4B (frozen)	MedGemma-4B + LoRA
LoRA rank / scaling	–	16 / 32
LoRA dropout	–	0.05
LoRA target modules	–	q, k, v, o, gate, up, down
EHR encoder d_{model}	192	frozen
EHR encoder layers / heads	3 / 4	frozen
EHR embedding d_{repr}	512	frozen
History decay λ (init)	1.0	frozen
Soft tokens per modality n_m	8	8
Projector hidden dim	1,024	1,024
Classification weight α	0.1	0.1
Memory bank k	–	3
Effective batch size	32 (micro 4 $\times$ accum 8)	2
Learning rate	2×10^{-4}	2×10^{-5}
LR schedule	Cosine (100 warm-up)	Cosine
Gradient clipping	$\ \mathbf{g}\ _2 \leq 1.0$	$\ \mathbf{g}\ _2 \leq 1.0$
Max sequence length	2,048 tokens	2,048 tokens
Mixed precision	BF16	BF16
Hardware	2 $\times$ H100 80 GB	2 $\times$ H100 80 GB

Table 9: Soft-token projector specifications. All projectors have a hidden dimension of 1,024 and produce an output shape of $8 \times 2,560$ (MedGemma-4B embedding size).

Modality stream	Input dim.	Stage
Current-visit ECG	d_{ecg}	1
Structured EHR	512	1
Longitudinal ECG history	512	1
Retrieved EHR context	512	2
Retrieved ECG context	d_{ecg}	2

D Per-Task AUROC: Deterioration, ICU, and Mortality

Table 11 reports per-task AUROC for all 15 non-diagnosis tasks on the held-out test set, together with 95% confidence intervals computed via empirical bootstrap (1,000 resamples per task). PRISM achieves AUROC above 0.95 on five tasks, including cardiac arrest, mechanical ventilation, and within-1-day mortality, and maintains AUROC above 0.90 on 11 of 15 tasks. Performance decreases as the mortality time horizon extends, consistent with the increasing contribution of chronic disease burden over longer intervals relative to the acute physiological state captured in the emergency department window.

E Diagnosis AUROC by ICD-10 Chapter

Table 12 aggregates AUROC across all 1,428 diagnosis tasks by ICD-10 clinical chapter under the generative pathway, sorted by mean AUROC. PRISM attains mean AUROC above 0.87 in 11 of 23 chapters and above

Table 10: Population memory bank configuration for PRISM.

Property	EHR Index	ECG Index
Vector dimension	512	d_{ecg}
Distance metric	Cosine similarity	Cosine similarity
Indexed encounters	116,186	116,186
Neighbors retrieved (k)	3	3
Context aggregation	Mean pooling	Mean pooling
Index construction	After Stage 1	After Stage 1

Table 11: Per-task AUROC for all Deterioration, ICU, and Mortality tasks on the held-out test set, with 95% bootstrap confidence intervals (1,000 resamples). The stronger pathway per task is **bolded**. Point estimates are computed on the retained sampled evaluation records; the bootstrap estimates uncertainty. The macro-averages in the final three rows are the exact means of the per-task values above them and reconcile with the corresponding entries of Table 1.

Group	Endpoint	Test n	PRISM _{disc} [95% CI]	PRISM _{gen} [95% CI]
Deterioration	Cardiac arrest	68	0.9706 [0.9187, 0.9965]	0.9870 [0.9627, 1.0000]
	Inotrope administration	27	0.9615 [0.8765, 1.0000]	0.9286 [0.8076, 1.0000]
	Mechanical ventilation	437	0.9597 [0.9438, 0.9759]	0.9666 [0.9517, 0.9789]
	ECMO cannulation	18	0.9383 [0.8051, 1.0000]	0.9383 [0.7916, 1.0000]
	Vasopressor initiation	125	0.9286 [0.8717, 0.9701]	0.9247 [0.8742, 0.9672]
	Severe hypoxemia	54	0.8285 [0.7069, 0.9293]	0.8285 [0.7129, 0.9286]
ICU	Admission within 24h	1,588	0.9191 [0.9061, 0.9329]	0.9221 [0.9100, 0.9349]
	Admission during stay	1,888	0.9118 [0.8987, 0.9239]	0.9136 [0.9002, 0.9249]
Mortality	During hospitalization	18	1.0000 [1.0000, 1.0000]	1.0000 [1.0000, 1.0000]
	Within 1 day	97	0.9864 [0.9574, 0.9983]	0.9817 [0.9540, 0.9979]
	Within 7 days	245	0.9404 [0.9101, 0.9669]	0.9421 [0.9086, 0.9680]
	Within 28 days	535	0.9188 [0.8939, 0.9414]	0.9224 [0.8993, 0.9442]
	Within 90 days	922	0.8970 [0.8762, 0.9150]	0.8971 [0.8772, 0.9158]
	Within 180 days	1,268	0.8759 [0.8573, 0.8950]	0.8761 [0.8561, 0.8928]
	Within 365 days	1,813	0.8771 [0.8625, 0.8915]	0.8780 [0.8629, 0.8926]
Deterioration macro-avg (6 tasks)			0.9312 [0.8958, 0.9541]	0.9290 [0.8916, 0.9582]
ICU macro-avg (2 tasks)			0.9155 [0.9070, 0.9247]	0.9179 [0.9092, 0.9265]
Mortality macro-avg (7 tasks)			0.9279 [0.9191, 0.9347]	0.9282 [0.9198, 0.9350]

0.80 in all chapters. The highest-performing chapters are those in which specific laboratory abnormalities are directly pathognomonic: obstetric conditions (Chapter O, mean 0.98), infectious diseases (Chapter A, 0.92), and circulatory disorders (Chapter I, 0.89). Performance is comparatively lower in chapters dominated by symptom-based or musculoskeletal diagnoses, where the primary discriminating information resides in clinical presentation rather than the structured biochemical record.

F AUROC Distribution Across Diagnosis Tasks

Table 13 stratifies all 1,428 diagnosis tasks by AUROC range under the discriminative pathway. The majority of tasks (78.4%) exceed AUROC 0.80, and more than one third (35.4%) exceed 0.90. Only eight tasks (0.6%)

Table 12: Diagnosis AUROC by ICD-10 chapter (PRISM_{gen}), sorted by mean AUROC. N : number of ICD-10 codes per chapter.

ICD-10 Chapter	Ch.	N	Mean	Range
Pregnancy, childbirth, puerperium	O	3	0.9825	0.9475–1.0000
Certain infectious diseases	A	15	0.9189	0.8493–0.9835
Blood and immune disorders	D	50	0.8952	0.6600–1.0000
Circulatory system	I	181	0.8900	0.7128–1.0000
Poisoning and toxic effects	T	43	0.8888	0.6317–0.9889
Neoplasms	C	48	0.8886	0.2800–1.0000
External causes (V)	V	38	0.8838	0.7376–0.9698
Factors influencing health	Z	166	0.8788	0.5091–1.0000
Infectious and parasitic (B)	B	37	0.8766	0.7381–0.9835
External causes (W)	W	11	0.8732	0.7784–0.9705
Endocrine, nutritional, metabolic	E	107	0.8703	0.6875–1.0000
Skin and subcutaneous	L	25	0.8681	0.6562–0.9835
Digestive system	K	119	0.8670	0.6942–1.0000
Genitourinary system	N	39	0.8660	0.6730–0.9952
External causes (Y)	Y	27	0.8577	0.7427–0.9917
Respiratory system	J	71	0.8519	0.5091–0.9905
Injury and trauma	S	43	0.8517	0.6317–0.9877
Nervous system	G	67	0.8431	0.6159–0.9738
Mental and behavioral	F	71	0.8411	0.5102–0.9982
Congenital malformations	Q	3	0.8411	0.7842–0.9224
Symptoms, signs, abnormal findings	R	171	0.8235	0.5091–0.9961
Eye and ear	H	24	0.8232	0.6311–0.9506
Musculoskeletal system	M	68	0.8038	0.4201–1.0000
Overall (1,428 tasks)		1,428	0.8624	0.2800–1.0000

fall below 0.60, all involving symptom-based or psychiatric diagnoses where the discriminating signal is largely absent from the structured laboratory and waveform record.

Table 13: Distribution of per-task AUROC across all 1,428 diagnosis tasks (PRISM_{disc}). Cumulative percentage is computed from the highest bin downward.

AUROC Range	Tasks	Proportion	Cumulative
≥ 0.90	506	35.4%	35.4%
0.80–0.89	613	42.9%	78.4%
0.70–0.79	260	18.2%	96.6%
0.60–0.69	41	2.9%	99.4%
< 0.60	8	0.6%	100%
Total	1,428	100%	

G Longitudinal ECG History: Availability and Aggregation

Table 14 reports how many prior ECG acquisitions each test encounter has within the 96-hour window. The distribution is strongly skewed, and it determines what the temporal aggregator can influence. For encounters with no prior acquisition the learned no-history embedding is used in place of $\mathbf{z}_{\text{hist}}$. For encounters with exactly one, the normalized weight of Equation 2 is 1 regardless of λ . The aggregation rule therefore alters the input for only the 122 encounters with two or more acquisitions, or 1.9% of the test fold, which bounds how far any two aggregation rules can differ on a macro-averaged metric.

Table 14: Prior ECG acquisitions per test encounter within the 96-hour window.

Prior ECGs	Encounters	Share
0	5,752	89.33%
1	565	8.77%
2	86	1.34%
3	26	0.40%
4	7	0.11%
5	2	0.03%
6	1	0.02%
≥ 2 (aggregator operative)	122	1.89%
Total	6,439	100%

Table 15 reports the converged decay parameter and the quantities derived from it. Because $|\lambda|$ remains within 1.1% of its initialization, the ratio $\tau/|\lambda|$ returns approximately τ by construction and should not be read as a timescale recovered from data.

Table 15: Converged temporal decay parameter and derived quantities.

Quantity	Value
Initialization of λ	1.0000
Converged $ \lambda $	0.9891
Change from initialization	1.1%
Half-life $\tau \ln 2/ \lambda $	67.3 h
$\tau/ \lambda $	97.1 h
Weight retained at 96 h	0.372
Weights at 0.4 h and 0.6 h	0.5005 / 0.4995

Table 16 compares the learned rule against fixed and uniform alternatives under both strata. The right-hand column restricts evaluation to the 122 encounters carrying two or more prior acquisitions, which is the only population on which the rules produce different inputs. The left-hand column covers the full test fold. Its narrow 0.09-point range is the arithmetic consequence of a 2.33-point effect confined to 1.9% of encounters, and it falls below the 0.32-point seed standard deviation of the full model.

H Cohort and Dataset Details

PRISM is trained and evaluated on the MDS-ED benchmark (Lopez Alcaraz et al., 2025), derived from the MIMIC-IV and MIMIC-IV-ECG databases (Johnson et al., 2023). This dataset constitutes the largest publicly available source of linked ECG waveforms and structured EHR data for emergency department encounters. We adopt the standardized 20-fold stratified splits defined by the MDS-ED protocol to ensure

Table 16: Alternative rules for pooling prior ECG embeddings.

Aggregation rule	All encounters	≥ 2 prior ECGs
Oldest acquisition only	0.9083	0.8903
Most recent acquisition only	0.9087	0.8998
Uniform average	0.9091	0.9026
Fixed decay, half-life 1 h	0.9090	0.9041
Fixed decay, half-life 2 h	0.9089	0.9069
Learned decay (PRISM)	0.9092	0.9136

direct comparability with all existing baselines. To represent the information available to a clinician at the point of disposition, all benchmark data is extracted from a 90-minute window beginning at ED arrival. Every encounter must have a raw 12-lead ECG, at least one structured laboratory measurement, and a complete discharge record with ICD-10-CM codes. The feature vector ($D=470$) includes demographics, biometrics, triage vitals, and laboratory panels; previous medications and free-text complaints are excluded. PRISM additionally consumes a retrospective 96-hour window ending at ED arrival, and the supplementary vital sign statistics of Section 3.2 and the prior ECG acquisitions of Section 3.4 are drawn from it. This window lies entirely before the benchmark window and therefore entirely before the prediction cutoff, so it introduces no post-cutoff information. It does give PRISM access to history that the MDS-ED and UniPACT baselines did not receive; the full-input benchmark comparison therefore includes this input-window difference.

Evaluation protocol. The test-set builder first retains all positives and an equal random sample of negatives for each task, yielding 265,816 exactly balanced encounter–task records. Records from 83 encounters with unavailable ECG waveforms are subsequently excluded: 81 encounters are absent from the MIMIC-IV-ECG index and two have unreadable waveforms. The retained evaluation contains 262,158 records and is not necessarily exactly balanced within each task. Initial class-balanced sampling follows the benchmark convention used by UniPACT (Tang et al., 2026); ECG-based exclusions are an additional selection step. AUROC measures positive–negative ranking, whereas average precision and calibration depend on the evaluation prevalence. We distinguish metrics on the retained sampled records from metrics adjusted to cohort prevalence. MDS-ED is a multi-label architecture rather than a per-task binary predictor, which is a structural difference between the systems and not a difference in evaluation. The benchmark test fold contains 6,454 task-eligible encounters, of which 6,439 have a resolvable prior-ECG history and are used in Appendix G, and 6,356 carry complete demographic coding and are used in Appendix I. Following the benchmark protocol, one ECG per stay is used. The Brier score and expected calibration error reported here describe calibration on the retained sampled evaluation records after ECG exclusions, rather than on an exactly balanced set or at natural cohort prevalence. They are therefore comparable across the systems and ablations reported here, which all use the same protocol, but should not be read as the calibration a deployed model would exhibit on an unselected ED population. Aggregate figures are computed as a macro-of-macro: per-task AUROC is averaged within each of the four task groups, and the four group means are then averaged. This prevents the 1,428 diagnosis tasks from dominating the fifteen deterioration, ICU and mortality endpoints, which are the outcomes of greatest clinical consequence. All reported systems use the same rule. Under task-weighted averaging the corresponding figures are 0.8626 for AUROC and 0.3426 for average precision at true prevalence.

The MDS-ED stratified folds are disjoint at the level of the patient, so no patient contributes encounters to more than one partition. The population memory bank is built only from training-fold encounters, so a validation or test query cannot return an encounter belonging to the same patient, and self-retrieval is precluded by construction. The same guarantee does not extend within the training fold, where a patient with multiple encounters may retrieve one of their own. Section 3.5 discusses the consequences. Table 17 summarizes the cohort statistics and experimental partitions. Table 18 describes the four clinical task groups predicted by PRISM.

Table 17: MDS-ED cohort and partition statistics.

Metric	Value
Unique patients	71,098
ECG-linked encounters	129,095
Training samples (task instances)	4.5M
Training set	116,186
Validation set	6,455
Test set	6,454

Table 18: Summary of the 1,443 binary clinical tasks predicted by PRISM.

Group	Tasks	Examples / Endpoints
Diagnosis	1,428	Full ICD-10 nosology (circulatory, musculoskeletal, etc.)
Deterioration	6	Intubation, cardiac arrest, vasopressors, severe hypoxemia
ICU Admission	2	Overall ICU admission, admission within 24 hours
Mortality	7	In-hospital; within 1, 7, 28, 90, 180, 365 days

I Performance by Patient Subgroup

Table 19 reports discrimination and calibration for every demographic subgroup recorded in the cohort, with 95% confidence intervals obtained by bootstrap over patients. Groups are ordered within each attribute by AUROC. Event rate is the proportion of positive instances in the retained sampled evaluation set for that subgroup, and differences in it reflect differences in case mix.

Table 19: PRISM performance stratified by sex, age band and self-reported ethnicity. Confidence intervals are bootstrap estimates over patients. Of the twenty-one pairwise contrasts within attributes, twenty have overlapping intervals; the exception is the 40–54 band against the 85+ band.

Attribute	Subgroup	Patients	Rows	Event rate	AUROC [95% CI]	ECE
Sex	Female	3,329	132,751	0.478	0.9177 [0.9037, 0.9270]	0.1143
	Male	3,027	129,407	0.523	0.8974 [0.8738, 0.9125]	0.1442
Age	<40	1,028	30,289	0.265	0.9277 [0.8867, 0.9481]	0.0962
	40–54	1,073	38,031	0.416	0.9240 [0.9054, 0.9363]	0.1153
	55–69	1,938	82,045	0.522	0.8969 [0.8745, 0.9141]	0.1378
	70–84	1,598	76,755	0.565	0.8868 [0.8602, 0.9064]	0.1356
	85+	719	35,038	0.599	0.8684 [0.8253, 0.8918]	0.1728
Ethnicity	Black / African American	1,391	53,856	0.457	0.9286 [0.9031, 0.9418]	0.1143
	Asian	225	8,267	0.473	0.9227 [0.8945, 0.9433]	0.1405
	Other	468	20,255	0.513	0.9153 [0.8831, 0.9369]	0.1253
	White	3,761	160,994	0.522	0.8917 [0.8733, 0.9061]	0.1379
	Hispanic / Latino	511	18,786	0.429	0.8222 [0.7735, 0.9229]	0.1290

Twenty of the twenty-one pairwise interval comparisons within attributes overlap; the exception is the 40–54 versus 85+ age bands. No comparison by sex or self-reported ethnicity separates. Interval overlap is descriptive and is not a formal test of equal performance. Excluding the Hispanic and Latino group discussed

below, subgroup AUROC is correlated with subgroup event rate at -0.78 . Across the age bands, event rate rises from 0.265 to 0.599 while AUROC falls from 0.9277 to 0.8684. This association is consistent with differences in case mix but does not establish that case mix causes the performance differences. Expected calibration error rises from 0.0962 in the youngest band to 0.1728 in the oldest under the sampled evaluation protocol with ECG exclusions, so the older bands also show larger calibration error in this analysis. The Hispanic and Latino subgroup has the lowest point estimate (0.8222) and a wide confidence interval spanning 0.1494. That interval contains the point estimate of every other ethnicity group, so the data do not support a reliable conclusion of differential performance. Its result also does not follow the same event-rate pattern as the other ethnicity strata. Whether this reflects a genuine difference or sampling imprecision cannot be settled here, and we flag it for further evaluation before deployment.

J Clinical Prompt Design and Representative Examples

Each training and inference instance is presented to the language model as a structured clinical question-answering prompt. The prompt opens with modality placeholder tokens, each replaced at forward time by an $8 \times 2,560$ soft-token matrix from the corresponding projector. It then provides a natural-language role description, a patient summary comprising demographic information, a compact time-series of vital sign measurements recorded in the 96-hour pre-arrival window, a stratified abnormality summary of all 470 structured clinical features, and a prior ECG history listing. The prompt closes with a binary prediction question instructing the model to answer with Yes or No. When a patient has no prior ECG acquisitions within the 96-hour window, the prompt contains four modality tokens. When prior recordings exist, a fifth token carrying the aggregated longitudinal ECG representation is inserted, resulting in a five-token prompt variant. The two boxes below illustrate representative prompts from the test set.

J.1 Box 1: Positive Case Without Prior ECG History

74-year-old male · 24-hour mortality · Ground truth: Yes

[EHR] [Retrieved EHR] [ECG] [Retrieved ECG] You are a critical care physician with access to the patient’s complete clinical record, current 12-lead ECG waveform, and Electronic Health Record (EHR). Your task is to predict whether a patient will experience mortality based on the ECG, the full 470-feature EHR summary, the vital signs time-series, and population evidence retrieved from similar patients. Patient: 74-year-old, Black/African, male.

Biometrics: not recorded. CRITICAL FINDINGS:

- Base excess: low at -20 mmol/L (normal -2.0 – 3.0 mmol/L).
- Glucose: critically elevated at 274 mg/dL (normal 70.0–100.0 mg/dL; $307 \rightarrow 274$ ↓, slope -0.107 /hr).
- INR: elevated at 1.3 (normal <1.1).
- Lactate: critically elevated at 10 mmol/L (normal <2.0 mmol/L; $9.9 \rightarrow 10$ ↑, slope $+0.0101$ /hr).
- pCO₂: critically elevated at 64 mmHg (normal 35.0–45.0 mmHg).
- PT: elevated at 15.1 sec (normal 11.0–13.5 sec).
- PTT: critically elevated at 60.5 sec (normal 25.0–35.0 sec).
- WBC: critically elevated at 15.6×10^3 / μ L (normal 4.5 – 11.0×10^3 / μ L). WITHIN NORMAL LIMITS:

Creatinine; Fibrinogen; Ionized calcium; Hematocrit; Hemoglobin; Platelets; RDW; RBC; BUN. Prior ECG History (within 96h):

No prior ECG recordings identified in the 96h window. Will the patient die within 24 hours? Answer strictly with Yes or No.

Box 2: Positive Case With Prior ECG History

80-year-old female · ESI Level 3 · Two prior ECGs · 24-hour mortality · Ground truth: Yes

[EHR] [Retrieved EHR] [ECG History] [ECG] [Retrieved ECG] You are a critical care physician with access to the patient’s complete clinical record, current 12-lead ECG waveform, and Electronic Health Record (EHR). Your task is to predict whether a patient will experience mortality based on the ECG, the full 470-feature EHR summary, the vital signs time-series, and population evidence retrieved from similar patients. Patient: 80-year-old, White, female. Acuity: ESI Level 3 (Urgent).

Biometrics: not recorded. CRITICAL FINDINGS:

- AST: elevated at 44 U/L (normal 10.0–40.0 U/L).
- Creatinine: critically elevated at 2 mg/dL (normal 0.6–1.2 mg/dL).
- Eosinophils: critically low at 0% (normal 1.0–4.0%).
- Glucose: elevated at 125 mg/dL (normal 70.0–100.0 mg/dL).
- Lactate: critically elevated at 5.1 mmol/L (normal <2.0 mmol/L).
- Lymphocytes: low at 14% (normal 20.0–40.0%).
- Neutrophils: elevated at 85% (normal 50.0–70.0%).
- Phosphate: elevated at 5.5 mg/dL (normal 2.5–4.5 mg/dL).
- Potassium: elevated at 5.2 mEq/L (normal 3.5–5.0 mEq/L).
- Troponin T: critically elevated at 0.11 ng/mL (normal <0.01 ng/mL).
- BUN: critically elevated at 39 mg/dL (normal 7.0–20.0 mg/dL). WITHIN NORMAL LIMITS:

ALT; Albumin; Alkaline phosphatase; Bands; Basophils; Bicarbonate; Total bilirubin; Total calcium; Chloride; Hematocrit; Hemoglobin; Magnesium; Platelets; RDW; RBC; Sodium; WBC. Prior ECG History (within 96h):

- 0.6h before current acquisition
- 0.4h before current acquisition Will the patient die within 24 hours? Answer strictly with Yes or No.

K Training Data Construction and Prompt Generation

Each training and evaluation instance presented to PRISM is a structured clinical question-answering record derived from the MDS-ED benchmark (Lopez Alcaraz et al., 2025). This appendix describes the end-to-end pipeline that transforms a raw MDS-ED task entry into the enriched prompt consumed by the MedGemma-4B backbone. The pipeline bridges the tabular MDS-ED data and the soft-token multimodal prompt format described in Sections 3.5 and 3.7; representative prompt instantiations are shown in Appendix J.

K.1 Source Data and Task Records

The 1,443 binary prediction tasks are sourced from the MDS-ED benchmark, which pairs each clinical encounter with an ordered set of question-and-answer records of the form “*Will the patient [outcome]? Answer: Yes / No*”. Each record contains a unique encounter identifier that maps to (i) a row in the MDS-ED feature CSV providing all 470 structured EHR features, (ii) a raw 12-lead ECG waveform path in MIMIC-IV-ECG, and (iii) a longitudinal patient entry in the 96-hour time-series store containing timestamped vital-sign measurements and a list of prior ECG acquisitions. No additional outcome annotation was performed; ground-truth binary labels are taken verbatim from the published benchmark to preserve strict comparability with all previously reported baselines on identical data partitions.

K.2 Population Statistics and Z-Score Normalization

Before any prompt is constructed, per-feature population statistics (mean and standard deviation) are computed over all numeric EHR columns from the 116,186 training-fold encounters (stratification folds 0–17 as defined by the MDS-ED protocol (Lopez Alcaraz et al., 2025)). These statistics serve two distinct roles. First, training-fold column medians are used for missing-value imputation in the FT-Transformer input vector (Section 3.2), preventing evaluation-fold leakage into the structured encoder. Second, they define the reference distribution for z-score-based abnormality flagging in the text prompt: a measurement is designated CRITICAL when it either falls outside the published clinical reference interval (Table 20) or yields a population z-score of magnitude greater than 2.0, whichever criterion is the more inclusive. Severity is further stratified as *elevated* or *low* when the fractional deviation from the nearest bound is at most 30%, and as *critically elevated* or *critically low* when the deviation exceeds 30%.

Table 20: Clinical reference intervals used for EHR abnormality flagging. Open bounds are indicated by a dash. Applied in conjunction with a population z-score threshold of $|z| > 2.0$ (Section K.2).

Cat.	Analyte	Lo	Hi	Unit	Cat.	Analyte	Lo	Hi	Unit
Vitals	Temperature	36.0	38.0	°C	Vitals	Systolic BP	90	140	mmHg
	Heart rate	60	100	bpm		Diastolic BP	60	90	mmHg
	Resp. rate	12	20	br/min		O ₂ sat.	95	–	%
Hem.	WBC	4.5	11.0	10 ³ /μL	Hem.	Neutrophils	50	70	%
	hemoglobin	12.0	17.5	g/dL		Lymphocytes	20	40	%
	hematocrit	36	50	%		Eosinophils	1.0	4.0	%
	Platelets	150	400	10 ³ /μL		Basophils	0	1.0	%
	RBC	3.8	5.8	M/μL		Bands	–	5.0	%
	RDW	11.5	14.5	%					
Coag.	INR	–	1.1	–	Coag.	PTT	25	35	sec
	PT	11.0	13.5	sec		Fibrinogen	200	400	mg/dL
Chem.	Sodium	136	145	mEq/L	Chem.	Potassium	3.5	5.0	mEq/L
	Chloride	98	106	mEq/L		Bicarbonate	22	29	mEq/L
	BUN	7.0	20	mg/dL		Creatinine	0.6	1.2	mg/dL
	Glucose	70	100	mg/dL		Total calcium	8.5	10.5	mg/dL
	Phosphate	2.5	4.5	mg/dL		Magnesium	1.7	2.2	mg/dL
	Albumin	3.4	5.4	g/dL					
Hepatic	ALT	7.0	56	U/L	Hepatic	AST	10	40	U/L
	Alk. phos.	44	147	U/L		Total bili.	–	1.2	mg/dL
Cardiac	Troponin T	–	0.01	ng/mL	Cardiac	Total CK	–	200	U/L
	CK-MB	–	5.0	ng/mL					
Metabolic	Lactate	–	2.0	mmol/L	Metabolic	pCO ₂	35	45	mmHg
	Base excess	–2	2.0	mmol/L		pH	7.35	7.45	–
	Ion. Ca	1.12	1.32	mmol/L					

K.3 Language Model Prompt

The prompt for each instance is assembled from five ordered components.

Modality token prefix. Between four and five placeholder tokens are prepended to the prompt, one per active modality: the current-visit ECG waveform (`<ecg>`), the FT-Transformer EHR encoding (`<ehr>`), the retrieved EHR population context (`<retrieved_ehr>`), the aggregated longitudinal ECG history (`<ecg_history>`, inserted only when the patient has at least one prior acquisition within the 96-hour window), and the retrieved ECG population context (`<retrieved_ecg>`). At forward time, each placeholder is replaced in-place by an $8 \times 2,560$ soft-token matrix from the corresponding projector (Table 9). Encounters without prior ECG recordings receive the four-token variant; all others receive the five-token variant.

Role and task block. A fixed natural-language role description identifies the model as a critical care physician and specifies the prediction target for the current instance (e.g., “*Will the patient be diagnosed with ‘Atrial fibrillation’?*”). This framing is derived verbatim from the binary question string in the MDS-ED task record and is standardized uniformly across all 1,443 tasks.

Patient summary block. The patient summary comprises four programmatically generated sub-sections: (i) a demographics and biometrics header containing age, sex, self-reported ethnicity, body-mass index, weight, height, and the Emergency Severity Index acuity level when recorded; (ii) a time-stamped vital-sign

table spanning up to ten representative rows from the 96-hour window (when more than ten measurements are available, the earliest and nine most recent observations are retained); (iii) a stratified EHR abnormality summary classifying all 470 features under CRITICAL FINDINGS or WITHIN NORMAL LIMITS according to the criteria in Section K.2, including observed value, unit, reference interval, and trend direction; and (iv) a listing of the number and relative timestamps (in hours before the index ECG) of prior ECG acquisitions within the 96-hour window.

K.4 Memory Bank Lookup Integration

After Stage 1 training, the FT-Transformer and D-BETA encoder produce fixed-length embeddings for all 116,186 training-fold encounters. These embeddings populate two independent nearest-neighbor indices, one per modality, as described in Appendix C. During the prompt-construction pass that precedes Stage 2 training and evaluation, each encounter is assigned its top- k nearest neighbors from each index via a single lookup against pre-computed neighbor tables materialized immediately after Stage 1. The resulting neighbor row indices are embedded directly in the serialized data record so that the Stage 2 data loader can reconstruct the corresponding embedding vectors at batch time without incurring query latency during distributed GPU training.

K.5 Train, Validation, and Test Handling

Validation and test encounters (stratification folds 18 and 19, respectively) are processed through an identical pipeline, with one critical exception: their memory bank queries target the *training-fold* index exclusively, ensuring that no ground-truth outcome information from held-out partitions enters the retrieved population context.

L Population Memory Bank: Retrieval Across Clinical Phenotypes

Figure 4 extends the single-encounter retrieval analysis of Section 4.4 to four clinically distinct test-set encounters, each representative of a different phenotypic category in the MDS-ED cohort. For each encounter, we take the 50 nearest neighbors of the EHR index, so all proportions below have a denominator of 50; ground-truth binary outcome labels and discharge diagnosis codes are aggregated by direct counting. The four selected cases illustrate different local neighborhoods. For example, the high-acuity metabolic case has neighbor prevalences of 30.0%, 40.0%, 26.0%, and 24.0% for mechanical ventilation, ICU admission, 28-day mortality, and cardiac arrest, respectively, while the chronic vascular case has 50.0% cardiac-arrest prevalence among its 50 EHR neighbors. The remaining cases show different diagnostic and outcome compositions. These examples are intentionally selected to span distinct phenotypes, and $k=50$ is used to make prevalence estimates readable. They illustrate index geometry only and do not establish cohort-level outcome concordance or the prediction mechanism.

M External and Temporal Evaluation Protocols

MC-MED cohort and prompts. Table 21 reports the temporal and external cohorts. The MC-MED cohort contains 40,121 visits from 30,642 patients. The local memory bank contains the 35,818 training and validation visits. Patient-overlap checks using medical record numbers found zero shared patients in all three partition pairs: training-validation, training-test, and validation-test. Consequently, the local memory bank, constructed from training and validation visits, contains no test patients. Training comprises 109,390 prompt records over 682 tasks, of which a stratified subset of 49,120 is used for fine-tuning. Validation contains 14,648 records over 314 tasks. The 13,254 test records cover 3,588 patients and 306 tasks; tasks enter the test prompts only if at least five positive examples are available.

MC-MED construction and local fine-tuning. Cohort construction reduces 118,385 source visits to 40,121 included visits. The available lead-II waveform is repeated across all 12 input channels for ECG encoding; this supplies a single measured lead rather than a diagnostic 12-lead recording. The structured-

Population Memory Bank: Multimodal Neighbour Retrieval Across Clinical Phenotypes

Metabolic Disorder Case — A
High-Acuity Gastrointestinal Profile

CLINICAL STATUS

Mech. ventilation 30.0% (15/50)	PRESENT
ICU admission (24h) 40.0% (20/50)	PRESENT
Vasopressors 14.0% (7/50)	ABSENT
28-day mortality 26.0% (13/50)	PRESENT
Cardiac arrest 24.0% (12/50)	PRESENT

NEIGHBOR DIAGNOSES (k=50)

C. Diff Enterocolitis	26%
Other Clinical History	22%
Equipment Dependency	20%

Endocrine Case — C
Endocrine Management

CLINICAL STATUS

Mech. ventilation 28.0% (14/50)	ABSENT
ICU admission (24h) 40.0% (20/50)	ABSENT
Vasopressors 28.0% (14/50)	PRESENT
28-day mortality 30.0% (15/50)	PRESENT
Cardiac arrest 46.0% (23/50)	PRESENT

NEIGHBOR DIAGNOSES (k=50)

Bacterial Infection	28%
Viral Gastropathy	26%
Enterocolitis	24%

Chronic Vascular Case — B
Chronic Vascular Case

CLINICAL STATUS

Mech. ventilation 20.0% (10/50)	ABSENT
ICU admission (24h) 36.0% (18/50)	PRESENT
Vasopressors 26.0% (13/50)	ABSENT
28-day mortality 26.0% (13/50)	PRESENT
Cardiac arrest 50.0% (25/50)	PRESENT

NEIGHBOR DIAGNOSES (k=50)

Bacterial Enteritis	26%
Enterocolitis	24%
Viral Gastropathy	24%

Vascular Risk Case — D
Vascular Risk Case

CLINICAL STATUS

Mech. ventilation 8.0% (4/50)	ABSENT
ICU admission (24h) 12.0% (6/50)	PRESENT
Vasopressors 0.0% (0/50)	ABSENT
28-day mortality 12.0% (6/50)	PRESENT
Cardiac arrest 28.0% (14/50)	PRESENT

NEIGHBOR DIAGNOSES (k=50)

Sepsis (Specified)	24%
Gram-negative Sepsis	14%
Intestinal Infection	14%

"Clinical Status" indicates ground truth. Values represent percentage and frequency (n/50) among the retrieved multimodal neighbors.

Figure 4: **Population memory bank retrieval across four clinical phenotypes (Cases A–D).** For each test encounter, the 50 nearest neighbors of the EHR index are used, so every percentage below is computed against a denominator of 50; clinical status (PRESENT/ABSENT) reflects the ground-truth outcome label of the index patient, and percentages report outcome prevalence among the 50 retrieved neighbors.

data mapping covers 43 of the 45 laboratory analytes. Local fine-tuning uses 49,120 stratified prompt records for two epochs. Training takes 2 hours 21 minutes on two H100 GPUs.

Eligible outcomes. Primary analyses retain tasks with at least five positive and five negative evaluation records. The temporal evaluation contains 841 tasks: 831 diagnosis tasks; mechanical ventilation and vasopressors; ICU admission within 24 hours and during the stay; and mortality within 1, 7, 28, 90, 180, and 365

Table 21: **Cohort sizes for external and temporal evaluation.** Counts describe the underlying visit/encounter tables, rather than the sampled encounter-task records. Banks comprise training and validation rows.

Setting	Partition	Encounters/visits	Patients
Temporal	Training	67,562	43,238
	Validation	3,524	2,342
	Test	1,435	1,002
	Memory bank	71,086	45,580
MC-MED	Training	31,239	22,908
	Validation	4,579	3,982
	Test	4,303	3,752
	Total cohort	40,121	30,642
	Local memory bank	35,818	26,890

days. The MC-MED evaluation contains 306 tasks: 297 diagnosis tasks; inotropes, mechanical ventilation, severe hypoxemia, and vasopressors; ICU admission within 24 hours; and in-hospital mortality and mortality within 1, 7, and 28 days. Group averages therefore summarize different eligible task sets across settings and are not direct estimates of a performance change on identical outcomes.

Prediction and aggregation. The PRISM rows in Tables 2 and 23 use the generative pathway: softmax over the Yes/No answer-token logits at the final prompt position. The discriminative pathway instead uses a classification head on the mean-pooled prompt representation. Baselines produce a single score. AUROC uses average-rank tie handling. MC-MED tasks are evaluated on positives and an equally sized random negative sample. For MDS-ED, balancing is performed before ECG exclusions over the full test fold; the temporal evaluation then retains later-era records without resampling. Neither ECG exclusion nor temporal restriction preserves exact within-task balance. All rows within each table use identical retained evaluation records. Per-task AUROCs are averaged within each eligible task group; overall AUROC is the unweighted mean of the four group means.

Uncertainty. Bracketed values are 95% percentile confidence intervals from 1,000 patient-level cluster-bootstrap resamples, using MIMIC-IV subject identifiers or MC-MED medical record numbers as the cluster unit. A common resample is used across compared runs, enabling paired contrasts. Fusion baselines use multiple seeds, whose scores are averaged within each resample. Each reported PRISM configuration uses a single training run. The gradient-boosting baseline reported in the MC-MED evaluation also uses a single run. These intervals characterize patient-sampling uncertainty and do not quantify PRISM’s training-seed variability.

MC-MED retrieval controls. Table 22 reports the group-specific uncertainty for the retrieval controls in Table 2. Both controls use the fine-tuned PRISM checkpoint and its generative pathway. Retrieval is perturbed at inference without retraining, with the evaluation records held fixed. The intervals describe each condition separately rather than the paired differences between conditions.

Source-table counts. The MDS-ED source tables used for the temporal evaluation contain 116,433 encounters from 63,929 patients in folds 0–17, 6,185 encounters from 3,525 patients in fold 18, and 6,439 encounters from 3,644 patients in fold 19. These source-table row counts are distinguished from the task-eligible counts reported for the original benchmark evaluation.

N Temporal Generalization

We additionally evaluate PRISM under a temporal split, training on encounters assigned with certainty to 2016 or earlier and testing on encounters assigned with certainty to 2017 or later. Encounters with date

Table 22: **MC-MED retrieval controls.** AUROC and 95% patient-level percentile-bootstrap confidence intervals for inference-time perturbations of fine-tuned PRISM. Overall is the unweighted mean of the four task-group AUROCs.

Task group	Retrieval zeroed	Random retrieval
Diagnosis	0.7920 [0.779, 0.807]	0.7919 [0.778, 0.806]
Deterioration	0.7389 [0.641, 0.831]	0.7310 [0.628, 0.828]
ICU	0.8681 [0.830, 0.903]	0.8657 [0.826, 0.901]
Mortality	0.9215 [0.887, 0.949]	0.9236 [0.889, 0.950]
Overall	0.8301 [0.802, 0.857]	0.8280 [0.799, 0.856]

Table 23: **Temporal generalization to later encounters.** All entries are AUROC; PRISM uses the generative pathway. Brackets show reported 95% patient-bootstrap confidence intervals. The later-era records are retained without rebalancing after the original test-set ECG exclusions. Zeroed and random retrieval are inference-time controls. Bold indicates the highest point estimate in each column, not statistical significance.

Model / condition	Diagnosis	Deterioration	ICU	Mortality	Overall AUROC
Fusion MLP	0.6617	0.8049	0.7858	0.8153	0.7669 [0.730, 0.801]
Fusion MLP + retrieval	0.6488	0.7941	0.7798	0.7935	0.7541 [0.716, 0.788]
PRISM	0.8119	0.8501	0.8726	0.9004	0.8587 [0.821, 0.892]
PRISM, retrieval zeroed	0.8064	0.8431	0.8665	0.8957	0.8529 [0.815, 0.887]
PRISM, random neighbors	0.8099	0.8525	0.8717	0.8988	0.8582 [0.821, 0.891]

intervals crossing the cutoff are excluded. Training and test patients are disjoint. Table 23 compares PRISM with compact fusion baselines on the same evaluation records. Overall AUROC is the unweighted mean of the four task-group AUROCs.

Partitions and evaluation coverage. Training comprises folds 0–17 restricted to encounters whose real year is certainly at most 2016; validation comprises fold 18 under the same restriction. Test encounters are from fold 19 with real year certainly at least 2017. The memory bank contains the union of training and validation, totaling 71,086 encounters from 45,580 patients. No test patient appears in the training, validation, or bank sets. Table 21 gives the partition counts.

The temporal evaluation retains the later-era records from the original sampled test set without resampling. It contains 59,493 encounter–task records from 998 of the 1,002 test patients. Eight of the 83 encounters excluded for unavailable ECGs are from 2017 or later and account for the four later-era patients with no retained record. Balancing was performed over the full test fold, so the later-era subset is not exactly balanced within tasks. The primary analysis includes 841 tasks meeting the five-positive/five-negative threshold; 1,270 tasks have a defined AUROC under the broader sensitivity criterion. Training contains 2,642,334 prompt records, with stratified subsets of 704,811 for Stage 1 and 99,820 for Stage 2. Validation contains 126,544 prompt records. PRISM uses the generative prediction pathway, and uncertainty follows Appendix M.

PRISM reaches an overall AUROC of 0.8587, compared with 0.7669 for the compact fusion MLP and 0.7541 for fusion with retrieval. These correspond to point-estimate differences of 9.18 and 10.46 AUROC points, respectively. PRISM also exceeds both compact fusion baselines in each task group. These comparisons concern the complete model configurations rather than isolating the language-model backbone.

Zeroing retrieval reduces PRISM’s overall AUROC by 0.58 points, but genuine retrieval exceeds random neighbors by only 0.05 points. These results distinguish sensitivity to removing retrieval context from the benefit of similarity-based neighbor selection: the latter is small in this temporal setting. Adding retrieval also does not improve the compact fusion baseline. We interpret these as setting-dependent effects and do not infer statistical significance from the separate model intervals.